RTO-MP-28
AC/323(IST)TP/4

RTO-MP-28

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25, 7 RUE ANCELLE,
F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

EUROPEAN SPEECH COMMUNICATION ASSOCIATION
c/o INSTITUT FÜR KOMMUNIKATIONSFORSCHUNG
UND PHONETIK
UNIVERSITÄT BONN, POPPLESDORFER ALLEE 47
D-53115 BONN, GERMANY

RTO MEETING PROCEEDINGS 28

Multi-Lingual Interoperability in Speech Technology

(l'Interopérabilité multilinguistique dans la technologie de la parole)

Papers and reports presented at the Tutorial and Workshop co-sponsored by the Information Systems Technology Panel (IST) of RTO-NATO and the European Speech Communication Association (ESCA) held in Leusden, The Netherlands on 13-14 September 1999.

Published August 2000

Distribution and Availability on Back Cover

RTO-MP-28
AC/323(IST)TP/4

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25, 7 RUE ANCELLE,
F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

EUROPEAN SPEECH COMMUNICATION ASSOCIATION
c/o INSTITUT FÜR KOMMUNIKATIONSFORSCHUNG
UND PHONETIK
UNIVERSITÄT BONN, POPPLESDORFER ALLEE 47
D-53115 BONN, GERMANY

RTO MEETING PROCEEDINGS 28

Multi-Lingual Interoperability in Speech Technology

(l'Interopérabilité multilinguistique dans la technologie de la parole)

Papers and reports presented at the Tutorial and Workshop co-sponsored by the Information Systems Technology Panel (IST) of RTO-NATO and the European Speech Communication Association (ESCA) held in Leusden, The Netherlands on 13-14 September 1999.

The Research and Technology Organization (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote cooperative research and information exchange. The objective is to support the development and effective use of national defence research and technology and to meet the military needs of the Alliance, to maintain a technological lead, and to provide advice to NATO and national decision makers. The RTO performs its mission with the support of an extensive network of national experts. It also ensures effective coordination with other NATO bodies involved in R&T activities.

RTO reports both to the Military Committee of NATO and to the Conference of National Armament Directors. It comprises a Research and Technology Board (RTB) as the highest level of national representation and the Research and Technology Agency (RTA), a dedicated staff with its headquarters in Neuilly, near Paris, France. In order to facilitate contacts with the military users and other NATO activities, a small part of the RTA staff is located in NATO Headquarters in Brussels. The Brussels staff also coordinates RTO's cooperation with nations in Middle and Eastern Europe, to which RTO attaches particular importance especially as working together in the field of research is one of the more promising areas of initial cooperation.

The total spectrum of R&T activities is covered by 7 Panels, dealing with:

• SAS	Studies, Analysis and Simulation
• SCI	Systems Concepts and Integration
• SET	Sensors and Electronics Technology
• IST	Information Systems Technology
• AVT	Applied Vehicle Technology
• HFM	Human Factors and Medicine
• MSG	Modelling and Simulation

These Panels are made up of national representatives as well as generally recognised 'world class' scientists. The Panels also provide a communication link to military users and other NATO bodies. RTO's scientific and technological work is carried out by Technical Teams, created for specific activities and with a specific duration. Such Technical Teams can organise workshops, symposia, field trials, lecture series and training courses. An important function of these Technical Teams is to ensure the continuity of the expert networks.

RTO builds upon earlier cooperation in defence research and technology as set-up under the Advisory Group for Aerospace Research and Development (AGARD) and the Defence Research Group (DRG). AGARD and the DRG share common roots in that they were both established at the initiative of Dr Theodore von Kármán, a leading aerospace scientist, who early on recognised the importance of scientific support for the Allied Armed Forces. RTO is capitalising on these common roots in order to provide the Alliance and the NATO nations with a strong scientific and technological basis that will guarantee a solid base for the future.

The content of this publication has been reproduced
directly from material supplied by RTO or the authors.

Published August 2000

ISBN 92-837-1044-4

Printed by Canada Communication Group Inc.
(A St. Joseph Corporation Company)
45 Sacré-Cœur Blvd., Hull (Québec), Canada K1A 0S7

Multi-Lingual Interoperability in Speech Technology

(RTO MP-28)

Executive Summary

Communications, command and control, intelligence, and training systems are more and more making use of speech technology components: i.e. speech coders, voice controlled C^2 systems, speaker and language recognition, and automated training suites. Interoperability of these systems is not a simple standardisation problem as the speech of each individual user is an uncontrolled variable such as non-native speakers using, additional to their own language, an official NATO language. For multinational operations, this may cause a reduced performance or even cause malfunction of an action. Standardised assessment methods and specifications for both commercial-off-the-shelf (COTS) and for development of new technology are required.

In the past the former DRG study group on speech technology (Panel-3, RSG.10) studied various effects of military environments in relation to the performance of speech technology focused on specific applications. Examples are the effect of noise on speech recognition, the effect of stress induced by workload, sleep deprivation, or battlefield stress, and presently the effect multi-linguality.

The present study considers interoperability of speech (communication) technology and embraces a wide range of military applications. It was identified that many nations, represented in the Task Group 001 of the IST-panel, have a major interest in command and control (speech recognition and synthesis), electronic war-fare (speaker and language recognition), training (communication operators, air traffic controllers) and understanding and translation systems.

In order to address these subjects a workshop on multilingual interoperability of speech technology was organised under responsibility of the RTO-IST-TG001 task group and the European Speech Communication Association. There were four tutorial papers and fifteen papers on a specific topic. The workshop took place in Leusden, The Netherlands from 13 to 14 September 1999. Over sixty people from twelve countries participated. Four topics were addressed in separate sessions:

Non-native speech and regional accents
Cross language speech processing
Identification of language and speaker
Human Perception and Assessment.

Each session was concluded with a plenary discussion. In these proceedings the tutorial papers, the topic related papers and a résumé of the discussions are given.

l'Interopérabilité multilinguistique dans la technologie de la parole

(RTO MP-28)

Synthèse

Les organismes C3, le renseignement et les systèmes d'entraînement font de plus en plus appel à des composants issus de la technologie vocale : il s'agit de codeurs vocaux, de systèmes C2 à commande vocale, de systèmes de reconnaissance du locuteur et du langage, ainsi que de programmes automatisés d'entraînement. L'interopérabilité de ces systèmes ne se présente pas comme un simple problème de normalisation, car la voix de chaque utilisateur individuel est une variable non-contrôlée, comme dans le cas d'un locuteur qui s'exprime dans une langue officielle de l'OTAN qui n'est pas la sienne. Dans le cas des opérations internationales, ce problème peut entraîner des performances réduites, voire même l'échec d'une action. Par conséquent, il y a lieu de définir des méthodes et des spécifications d'évaluation normalisées, tant pour les produits du commerce (COTS), que pour le développement de nouvelles technologies.

Dans le passé, le groupe d'étude sur la technologie vocale de l'ancien GRD (Panel-3, RSG.10), a examiné les différents effets des environnements militaires sur les performances de la technologie vocale pour des applications spécifiques. Des exemples de telles applications sont les effets du bruit sur la reconnaissance vocale, l'effet du stress engendré par une surcharge de travail, le manque de sommeil, le stress du champ de bataille, et récemment, l'effet multilingue.

Cette étude examine l'interopérabilité des technologies vocales (communication) et couvre un large éventail d'applications militaires. Il a été constaté que de nombreux pays représentés au Groupe de travail 001 de la commission IST s'intéressent vivement au commandement et contrôle (reconnaissance et synthèse de la parole), à la guerre électronique (reconnaissance du locuteur et du langage), à l'entraînement (opérateurs de communications, contrôleurs de la circulation aérienne) et au système d'analyse et de traduction.

Afin d'examiner ces sujets, un atelier sur l'interopérabilité multilingue de la technologie vocale a été organisé sous l'égide conjointe du groupe RTO-IST-TG001 et de l'Association européenne de la communication vocale. En tout, quatre communications pédagogiques et quinze communications spécialisées ont été présentées. L'atelier a été organisé à Leusden, aux Pays-Bas, les 13 et 14 septembre 1999. Plus de soixante personnes, de douze pays différents y ont participé. Quatre sujets ont été examinés lors de quatre sessions distinctes, à savoir :

Les locuteurs non-natifs et les accents régionaux
Le traitement de la parole interlingue
L'identification du locuteur et du langage
La perception humaine et l'évaluation

Chaque session a conclu par une discussion plénière. Ce compte rendu de conférence contient les communications et un résumé des discussions.

Contents

SESSION "NON-NATIVE SPEECH AND ACCENTS"
CHAIRED BY DENIS JOHNSTON, BT, ENGLAND

SESSION "HUMAN PERCEPTION AND ASSESSMENT"
CHAIRED BY EDOUARD GEOFFROIS, DGA, FRANCE

SESSION “CROSS LANGUAGE” CHAIRED BY TIM ANDERSON, WPAFB, USA

SESSION “IDENTIFICATION” CHAIRED BY LOUIS BOVES,
UNIVERSITY OF NIJMEGEN, THE NETHERLANDS

Foreword

Interoperability of systems is of crucial importance. When speech and language come into play, interoperability of systems developed for specific languages becomes an issue. Several different situations must be envisaged. For instance, one might want to use a speech coder optimised for American English in German or French. Or a native speaker of Dutch might want to use a speech recogniser trained for Spanish. These examples show that interoperability is an important issue for many applications of modern speech technology. For this reason a special task group of the NATO Research and Technology Organisation started a project on the development and assessment of multi-lingual applications of speech coding, speech recognition, topic spotting, speaker and language identification, and speech synthesis.

In the past this task group organised a number of workshops in co-operation with ESCA thus initiating an interaction between civil and military applications that to a large extent pose the same requirements. Recent workshops based on this concept were: "Applications of Speech Technology", Lautrach-Germany 1993, "Speech under Stress", Lisbon-Portugal 1995, and "Robust Speech Recognition for Unknown Communication Channels" Pont-à-Mousson-France 1997.

The program of the "MIST" workshop covers four themes:
Non-native speech and regional accents
Cross language speech processing
Identification of language and speaker
Human Perception and Assessment.

Four tutorial lectures introduce the various sessions of the workshop. Additionally, each session concluded with a plenary discussion. A résumé of these discussions are included in these final proceedings.

I would like to thank the NATO-RTO and ESCA for their support in the organisation of the workshop; the tutorial speakers, the discussion leaders and reporters for their time, effort and expertise; the International Scientific Committee for their help in reviewing the proposals and their constructive advise. Finally, I would like to thank my colleagues of the local organising committee who spent a lot of their time supported by their enthusiasm to ensure the very promising programme, for taking care of all of the logistics and for editing these final proceedings.

Herman J.M. Steeneken

Committee Members

Organising Committee

Herman J.M. Steeneken

Elisabeth den Os

Sander J. van Wijngaarden

David A. van Leeuwen

Johan Koolwaaij

International Scientific Committee

Roberto Billi	(Italy)
Hervé Bourlard	(Switzerland)
Louis Boves	(Netherlands)
Dirk van Compernolle	(Belgium)
Sadaoki Furui	(Japan)
Melvyn Hunt	(UK)
Denis Johnston	(UK)
Lori Lamel	(France)
Tony Robinson	(UK)
Christelle Sorin	(France)
Isabel Trancoso	(Portugal)
Marc Zissman	(USA)

SPEECH RECOGNITION BY GOATS, WOLVES, SHEEP and ... NON-NATIVES

Dirk Van Compernolle
Lernout & Hauspie Speech Products NV
Koning Albert I Laan 64, 1780 Wemmel, Belgium
Tel. +32 2 456 05 00, Fax +32 2 460 01 72, E-mail Dirk.VanCompernolle@lhs.be

ABSTRACT

This paper gives an overview of current understanding of acoustic-phonetic issues arising when trying to recognize speech from non-native speakers. Regional accents can be modeled by systematic shifts in pronunciation. These can often better be represented by multiple models, than by pronunciation variants in the dictionary. The problem of non-native speech is much more difficult because it is influenced both by native and spoken language, making a multi-model approach inappropriate. It is also characterized by a much higher speaker variability due to different levels of proficiency. A few language-pair specific rules describing the prototyical nativised pronunciation was found to be useful both in general speech recognition as in dedicated applications. However, due to the nature of the errors and the mappings, non-native speech recognition will remain inherently much harder. Moreover, the trend in speech recognition towards more detailed modeling is counterproductive for the recognition of non-natives.

INTRODUCTION

That recognition of non-native speech is significantly harder than that of native speech can't be a surprise. We as humans often have a hard time understanding someone speaking his second or third language. We might also readily determine the accent and will quickly make an assessment on the degree of non-nativeness.

But we also know that there is not something like "a non-native". French, Japanese and Indians will speak English in a very different way. The sounds will not just be accented, but they will insert and delete phonemes, they will make grammatically weird sentences, etc. After some time we may get used to the peculiarities of their speech and understand them quite well. Listening to another non-native in a language, non-native for ourselves, sometimes turns out not to be too difficult because the speaker uses a restricted vocabulary and easy syntax.

A speech recognizer is often compared to a person who is bad of hearing, a young child or to someone who isn't too familiar with the language. So maybe a recognizer should like non-native speech. We'll see that this is not at all the case as the recognizer will take little or no advantage from the reduced grammatical complexity, but will suffer greatly under miserable acoustic phonetic conditions. So a recognizer will only see the bad sides of non-native speech and generally poor robustness of speech recognition systems will show double.

In this review paper I will focus on the acoustic phonetic issues. It is structured as follows. First I'll discuss native accents; then I will revisit the complexity of differences in phoneme spaces across languages, moving on to the complexity of non-native speech recognition for general purposes and dedicated applications.

ACCENTS AND DIALECTS

CHARACTERIZING ACCENTS

Each living language has numerous accents which are continuously on the move. It's sometimes implicitly assumed accents will differ most distinctively in the realization of vowels[Bary89], but consonantal differences may be strong as well. Eg regional distinctions in Latin American Spanish are especially pronounced for a few consonants.

Accents will only show minor differences at the higher - abstract - phonemic level, but the specific acoustic-phonetic realizations might shift considerably. Small phonetic shifts can freely be applied to almost all sounds of any language without having any impact on recognition as all languages only use a limited part of the articulatory space. As phonemic ambiguity shouldn't increase markedly by accent shifts, a strong shift of one class could have a forceable impact on other classes as well. It is possible that accents introduce or remove homonym confusions, but overall acoustic confusability should not change significantly.

In terms of pattern recognition one might describe an accent as a shift in classes across the feature space, but with maintenance of the same degree of separability of the classes. Typical of native accents is that these shifts will be applied in a pretty consistent manner by whole groups of speakers.

There have been two main paths in attacking the dialect problem for speech recognition. The first one tries to to model accents as pronunciation variants at the detailed phonetic level[Bary89,Cohe89,Adda98]; the other one doesn't get involved with detailed modeling but creates multiple models for large speaker groups [VCom91,Beat95,Drax97].

Existence of accents questions the validity and feasibility of symbolic representation of sounds, but at the same time highlights the tremendous abstraction applied in our alphabetic writing systems. At the abstract (phonological) level a unique symbolic representation may suffice for a whole group of accents. If, on the contrary, we want to represent all the different realizations in a symbolic (phonetic) way, then the better chance is that no system will be detailed enough. Straightforward reasoning also leads to a few more conclusions. Because of the continuity of the shifts that are feasible at the pronunciation level, any symbolic representation is inherently local and not universal. Abstraction and symbolic representation are hence not absolute but relative and only valid within the applicable language. Phoneme boundaries aren't absolute, but defined wrt. to the collection of phonemes valid for that language. Ultimately it follows directly from the continuity of the characteristic sound shifts, that granularity and categorization of dialects is a very ill-defined problem.

Now, let's confront the above hypotheses with experiences with real world speech recognition. The Dutch/Flemish language group is an interesting case study as accent and dialect diversity is tremendous, given its compact geography, but we'll restrict to the case where everyone at least attempts to speak the "standard" language and not the local dialect. Contrary to the British/US English distinction there are no spelling differences between Dutch and Flemish.

MODELING ACCENTS BY MULTIPLE ACOUSTIC MODELS

Everyone who has tackled the problem of Dutch/Flemish speech recognition knows that models trained on one group perform very poor on the other group. Error rates may double or triple. Relaxing within class variability will not help, because it isn't random extra variability that needs to be modeled, but a systematic shift. Putting all data in a single model gives reasonably satisfying results, but will still be significantly (eg 20%) worse than accent specific models. There are also some interesting asymmetries showing increased or decreased separability for certain classes depending on the accent. One such example are the digits. For Dutch speakers the pair 'twee/twe:/-drie/dri/' (similar as for German zwei-drei), while for the Flemish the pair 'vijf /vɛif/ - zes/zes/' is by far the more confusable one. The above can be understood by following two characteristic differences of Dutch vs Flemish:

- Diphtongization of long vowels by Dutch, reduces the ee-ie phonetic distances. This goes together with a stronger diphtongization of the real diphtongs in Dutch vs Flemish which increases the distance of ei-e
- Devoicing of voiced fricatives, which is stronger however for the /v/ than for the /z/ which increases the phonetic distances of the v-z pair.

Interesting to note is that the above shifts get more pronounced the further north one goes and that the geographical boundary for these phenomena might even be better characterized by the Maas-Rijn Delta than the Belgian-Dutch border.

The strength of the shifts - up to the phonemic level - causes a strong overlap of distributions in a global modal, while accent specific distributions are much better separated. The latter may be a good criterion to decide if accents should be modeled as extra speaker variability in a single model or if multiple models are required. The above is also a good example that accent shifts can either somewhat reduce or enhance phonetic contrasts between words. These small changes may have little impact on human performance, but show up in machine based recognition.

Now that usage of 2 models for Dutch/Flemish seems perfectly reasonable, one may wonder how many more models would make sense and how to define them. In some early work on this problem [VCom91] it was found that extra models based on regional clustering provided little or no advantage, but the interpretation may have been influenced by insufficient data to train a larger number of models. In unrelated more recent work, it was found that 3-4 models does make sense.

In similar experiments for US English [Beat95], it was found that 3 accent models for the US gave a good tradeoff between performance, compactness and trainability of the models.

Overall we can conclude that using multiple models for the different dialects is an easy and effective way to improve performance. Modeling of a very small number of well designed large clusters seems to perform better than many small clusters, because of loss in intrinsic speaker variability in the clusters when insufficient training data is available.

MODELING ACCENTS BY PRONUNCIATION VARIANTS

The strong phonetic differences between Flemish and Dutch or British and US English would intuitively suggest another way to model strong accent differences, i.e. by pronunciation variants[Cohe89]. In last year's

ESCA workshop on pronunciation variation much interesting work was presented [eg Adda98,Rile98], but often with somewhat disappointing results. Only the most pronounced variants are essential, especially so for the most frequent short words of a language. When modeling variants in great detail, eg for speaking style differences, then increased confusability seems to offset the increased modeling capacity.

A major weakness of implementing accent variability by multiple pronunciations in a single dictionary is that accent consistency for a given speaker is not enforced. Therefore, another approach - which is rarely feasible in real-time speaker independent systems - is the use of parallel phonetic dictionaries, with dictionary selection on a maximum likelihood criterion. This is easily done however in speaker dependent and/or speaker adaptive dictation systems where the choice can be based explicit speaker preference or after parallel batch processing.

In the speech recognition world British and US English are most often treated as 2 different languages with different spellings, separate phonetic dictionaries - probably even different phonetic alphabets. It comes somewhat more intuitive than in the Flemish/Dutch case because of the spelling differences and the geographical separation. Nevertheless, it can be shown that speech recognition performance will still be very reasonable if the phonetic baseforms from one variant are used for the other, but trained with the correct speaker group. It shows great resilience of phonetic transcriptions against accent variation as long as the canonical transcription only needs to be valid for a coherent regional group and not for multiple groups at the same time. This is explained by the fact that most pronunciation variants will be learned implicitly when building context dependent acoustic models.

CROSS-LINGUAL PHONETICS

IPA AND ITS COMPUTER EQUIVALENTS

Alphabetic writing systems must stand out as one of the greatest inventions of all times. It made it possible to write about every language with as few as 30 symbols, corresponding to the sounds of the language. Due to independent evolution of the Roman alphabet in different languages and further emphasized by the independent evolution of written and spoken language, the phonetic consistency is far away in most of today's languages and complicated grapheme-2-phoneme converters are necessary to go from written to spoken language.

Modern phonetic alphabets are in a way a reinvention of the original alphabet and try to write according to the rule "one sound - one symbol ". The IPA (International Phonetic Alphabet) is the concerted international effort that tries to achieve this (illusive) goal for all languages of the world at once. That each language only sparsely fills the articulatory and acoustic space is well illustrated by the fact that the IPA needs several hundred basic symbols to encompass all languages. Several ASCII compatible computer derivatives are used by speech community has derived its own derivatives (SAMPA, Worldbet). At L&H we developed our own version L&H+ for internal usage. These cross-lingual phonetic alphabets greatly enhance readability but at the same time create the false impression (hope) of the existence of a truly language independent phonetic alphabet.

Extensive experience over the past 5 years in speech technology applications has shown how illusive the target "one sound - one symbol" might be. L&H+ foresees in about 300 different classes for the 30 odd languages that it is currently used for. Despite all efforts and good definitions, there remains a great lack of inconsistency between transcriptions in different languages. This is due to the enforcement of a single symbol on multiple classes which are close but not truly identical. One of the complicating aspects is that no phonetician exists who can claim native or close to native pronunciation for a sufficiently large group of languages. Thus even the best implementation is based on a consensus of experts who don't really understand each other.

LISTENING AND SPEAKING BY NON-NATIVES

There are many similarities but also a few significant differences between accents of natives and pronunciations of non-natives. Class definitions are only valid within a single language (and accent) and there is no reason whatsoever why class definitions of one set should be portable to another one. The very fine distinctions will get lost in any compact symbolic representation. Similarly some of those distinctions we do hear and others we don't. Which distinctions we hear, depends much on our language exposure at younger age. It's not so extreme that we have learned strict class boundaries applicable only to our native language, but it seems that we have learned to listen for sound features which are most relevant to our native language[Fox95], somehow projecting all acoustic features onto a lower dimensional space appropriate for our native tongue. And by feedback mechanisms our acoustic and articulatory spaces are tightly coupled, so we only pronounce those sounds adequately that we need in our native tongue.

Numerous straightforward examples can be given. The tonal phonetic features of oriental languages are tough to hear and learn for Europeans because it didn't get engraved in their front end acoustic processor. Somewhat less pronounced, but well demonstrated, is that natives of different European languages might discriminate vowels along different feature dimensions [Fox95]. Thus what is a phonemic distinguishing feature for a native of one language may hardly be audible to a native of another one. Consequently you must expect that a non-native will significantly mispronounce sounds that are not in his native auditory collection, by projecting the pronunciation onto his own articulatory and acoustic space. As an example, don't be surprised if you hear a Spanish person mention *'a shit of paper'*, by omission of the duration cue in the word *sheet.* Similarly, I shouldn't be too surprised if both human and machine recognizers mistakes my 'p' for a 'b' by lack of aspiration of the 'p'. While the aspiration is a distinguishing feature in English it is not in Flemish, where it does not exist.

Thus there are significant differences between native and non-native accents. Native accents are all based on pretty much the same phoneme set. Because of proximity, it is reasonable to assume that acoustic feature space and distinguishing acoustic clues will be very similar and the average phonemic contrast will be maintained across native accents. Native accents are information preserving transformations. Non-natives will project sounds onto a subspace defined by the intersection of target language and native language, thus on an inherently lower dimensional feature space, thus potentially with loss of information. And the further that languages are apart from each other, the worse the intersection will be and the greater the information loss [Bona98].

MULTI-LINGUAL SPEECH RECOGNITION

Our inherent skepticism about cross lingual phonetic alphabets can be put to test by a multi-lingual speech recognition system.

In recent years, several groups have tried to build cross language phone models. The ultimate goal would be that one sufficiently large collection of phoneme models is sufficient to model all the languages of the world. But more often the goals are more restrictive. It is either used to have a compact footprint for multilingual systems or to bootstrap or augment the training of acoustic models in a new language when little data is available [Köhl96,Bona97,Schul98].

At L&H we've also used such systems to deal with initial responses in a multi-lingual system with a priori unknown language by the caller. This avoids the problem of ranking scores between 2 systems with completely different models. The results we found are similar to the ones found elsewhere in the literature.

- Multilingual phone models perform worse than single language phone models, provided there is enough training data for each of the languages
- The effect becomes more pronounced as more diverse languages are grouped together. This is naturally explained on the basis that phoneme classes from far away languages cluster intrinsically less good, but it may also be a hint that the multilingual phonetic alphabet misses some important details.
- Degradation may be on the order of 20-80% depending on the number and diversity of languages that are clustered.
- Despite their poorer performance, such systems may have a high practical value, especially when little or no data exists in a particular language or in some simple but intrinsic multilingual tasks

NON NATIVE SPEECH RECOGNITION

MORE DATA OR DIFFERENT MODELS ?

Based on the above, the easy way out might be to consider non-native speech as just another (heavy) accent. If the occasional pronunciation errors are modeled as random then we can even forget about them. So all we need is data. To some extent it is a valid approach, except that ... variability is much larger and non-natives are by no means a homogeneous group. At least the influence of the native language needs to be taken into account. Thus, if we need to start collecting data on non-natives, then the whole data collection problem becomes quadratic in nature and is clearly not feasible nor can it be the right approach. Here we are just running into the limits of more and more data. Assuming that the data problem is quadratic might even be underestimating the real dimensionality. It is well known that people talking in their third, fourth .. language might copy - correctly and incorrectly - pronunciations from other foreign languages they know. All of this is further complicated by the large variability in language proficiency among the non-natives.

So is there anything else to do than lay back and observe that non-natives are worse than natives ? Digging deeper, the situation looks even more grim. Much of the progress in the last 15 years in acoustic modeling is based on more detailed modeling, by creating sharper and sharper distributions for narrower and narrower classes. This is diametrically opposite of the tolerance and robustness required for non-natives. Distribution of non-native scores on allophonic variants will greatly differ from the distribution of natives, because they will emphasize different cues. So it should come as no surprise that for people with heavy accents the performance gain between context-independent and context-dependent models might get totally washed out.

SPEAKER ADAPTATION FOR NON-NATIVES

There is another feature about non-natives which has significant impact on ASR systems. Vocabulary of non-natives tends to be much more limited and occurrence of unknown words will not be uncommon. These are likely to happen in enrollment scripts. Whenever an unknown word occurs, the speaker will hesitate and apply certain letter-2-sound rules, typically a mix of the rules of his native tongue mixed with the non-native one, leading to all kind of funny pronunciations.

Potential for speaker adaptation will thus greatly depend on proficiency of the non-native. If all words in the adaptation script are known to the non-native, then we fall back to the 'thick accent' case. If there are many unknown words, hesitations will occur and gross mismatches between pronunciation and transcription will be present. Such mismatches will not shift the sound categories to their desired location, but will randomly smear out the distributions. One way to avoid this is to include only speech with minimal confidence levels, but as could be expected, this is even more difficult for non-natives. For reasonably proficient non-natives, speaker adaptation has shown dramatic improvements[Zava95] reducing the error rate by a factor 2-3 without adaptation of the phonetic baseforms. This confirms the assumption that a very strong accent shift needs and can be modeled by transformation of the distributions. However, even after adaptation, non-natives still performed a factor 2 worse than natives. This is explained by a combination of effects: (i) random pronunciation errors and (ii) projection of pronunciation onto a lower dimensional, less discriminative, space. Another more subtle cause may be that the chosen state tying - necessary in speaker adaptation - is optimized for natives and might be less applicable to the non-native accents.

NATIVISED PRONUNCIATIONS

Pronunciation errors are common with unknown words, and even more so if simple letter-2-sound rules are insufficient as is the case for proper names - a common problem in Europe with its density of languages and high mobility. The two most immediate application areas are automated attendants and car navigation.

The automated attendant in our office is a good example of how complex a small problem quickly gets. There are roughly 100 employees of whom about 60% are Flemish natives of whom most but not all have a name with Flemish pronunciation. The only other significant language group are the French speakers. In total there are names of 12 different language origins of which 4 from outside Europe. Despite the monolingual English greeting, the name pronunciation might be in many different ways, given in order of occurence: native pronunciation, pronunciation with a Flemish accent, pronunciation with an English accent, pronunciation with another accent. This is in stark contrast to the implementation of similar systems in US or France, where almost all users would have a tendency to bastardize the name pronunciations to the local language. Given the great mix of pronunciation and accent, there is no option for a language-pair specific solution and one needs to rely on some "language independent" recognizer as the symbol set from a single language will be insufficient to code all the various transcriptions that one might require. On average 2-3 transcriptions of each name suffice to yield acceptable performance. Given the sparseness of the language mix, we did not make great attempts to derive general rules that would describe prototypical pronunciation variants. The system has been operational internal for several years and many of us have learned fail safe pronunciations for the names we often use.

Another case is the one of car navigation, as explored in the EC VODIS project[VODIS]. Assume a German travelling to France and talking to the navigation unit in German while specifying French location names.

It was found that Germans - also the ones with little French knowledge - have some knowledge of French phonology and ultimately use a mix of French and German letter-2-sound rules[Tran99]. A reasonable approximation of the real pronunciations is obtained by starting from the correct French pronunciation and applying a small set of French-2-German conversion rules. Most of these can be related to the absence of a very close relative of a particular sound in the native language.

While done in an ad hoc manual way, part of the above work can be automated and common mutations could be learned on the basis of a moderate body of German pronunciation of French names. At the same time it becomes obvious that many similar rules - but maybe somewhat reduced - would apply for a German speaking French. Similar rule based work has been reported in the field of nonnative speech recognition [Bona98], pronunciation variation in general [Crem98]. Today, this may stand out as one of the more promising approaches in dealing with non-native pronunciations.

LANGUAGE LEARNING

One of the most extensively researched topics in non-native speech recognition is the one of language learning[Stil98]. For this application there may be many more novice speakers than others who have already a thorough knowledge. The most intuitive measure to evaluate someone's pronunciation is some form of confidence measure. But similarly as with native speech recognition, simple likelihood measures aren't a most reliable metric, and it's correlation with expert ratings was found to be low[Neum96]. It was found that rate-of-speech [Cucc98,Neum96] is a reliable estimator of degree of non-nativeness. However, ROS has little diagnostic value as it does not identify pronunciation errors.

Likelihood scores can be turned into a much more reliable measure if they are turned into a likelihood ratio of speaker vs. prototypical native pronunciation. In order to obtain a reference score pronunciations of 10-20 native speakers of all sentences in a lesson can be recorded and processed by the recognizer. This procedure has been found to yield significantly better performance than the use of more generic methods to generate the reference score in the likelihood ratio.

Still an alternative approach for turning likelihood ratios into indicators of pronunciation errors, is the explicit modeling of expected errors. Due to the very different phonotactic structure of Japanese vs. English, many pronunciation errors made by Japanese, learning English, can be predicted[Kawa98]. Consonant clusters, which are non existent, will lead to vowel insertions and diphtongs are likely to be replaced by a single vowel. A pronunciation network including the correct and incorrect pronunciations is subsequently fed to the recognizer and simple Viterbi alignment shows immediately all errors. The latter approach is very efficient for the small group of frequent language-pair specific errors. Basically the same set of rules applies as discussed in the previous section on nativised pronunciation.

CONCLUSIONS

In this paper we reviewed the difficulties arising when recognizing non-native speech, especially the additional difficulties compared to dealing with native accents. Non-natives are more complex than heavy accented speakers. Across different applications it was found that a few language pair specific rules can describe many of the typical mispronunciations. However, because the loss of certain distinguishing acoustic cues and heavy shifts in pronunciation, non-native recognition will be very difficult for today's recognizers using sharp distributions. It stresses the inherent lack of robustness of our current acoustic-phonetic modeling. Likelihood scores should gracefully decay as phonetic feature distance grows which is not necessarily the case in a state of the art recognizer.

ACKNOWLEDGEMENTS

Much of the discussion presented in this paper is based on unpublished work at L&H from Jose Conejo, Kristin Daneels, Bart D'Hoore, Luc Mortier, Louis ten Bosch, and Filiep Vanpoucke.

REFERENCES

[Adda98] M. Adda-Decker, L. Lamel, "Pronunciation variants across systems, languages and speaking styles", Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, pp.1-6, Rolduc, May 1998.

[Bary89] W.J. Bary, C.E. Hoequist and F.J. Nolan, "An approach to the problem of regional accent in automatic speech recognition", Computer Speech and Language, 3, pp.355-366, 1989.

[Beat95] V. Beattie et. al. "An integrated multi-dialect speech recognition system with optional speaker adaptation", Proc. Eurospeech95, pp.1123-1126.

[Bona97] P. Bonaventura, F. Gallocchio, G. Micca, “Multilingual speech recognition for flexible vocabularies”. Proc. Eurospeech’97, pp 355-358, 1997

[Bona98] P. Bonaventura, F. Gallocchio, J. Mari, G. Micca, "Speech recognition methods for non-native pronunciation variants", Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, pp. 17-22, Rolduc, May 1998.

[Cohe89] M. Cohen "Phonological Structures for Speech Recognition", Ph.D. Thesis, UC Berkeley, 1989.

[Crem98] N. Cremelie, J.P. Martens "In search for pronunciation rules", Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, pp. 23-28, Rolduc, May 1998.

[Cucc98] C. Cucchiarini, F. de Wet, H. Strik and L. Boves "Assessment of Dutch pronunciation by means of automatic speech recognition technology", Proc. ICSLP 98, pp.751-754.

[Drax97] C. Draxler and S. Burger, "Identification of regional variants of high German from digit sequences in German telephone speech", Eurospeech 97, pp.747-750.

[Fox95] R.A. Fox, J.E. Flege and M.J. Munro, "The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis". JASA Vol.97(4), pp. 2540-2511, 1995.

[Kawa98] G. Kawai, K. Hirose "A method for measuring the intelligibility and Nonnativeness of phone quality in foreign language pronunciation training", Proc. ICSLP 98, pp.782-785.

[Köhl96] J. Köhler "Multilingual phoneme recognition exploiting acousitc-phonetic similarities of sounds" Proc. ICSLP96, pp.2195-2198.

[Neum96] L. Neumeyer, H. Franco, M. Weintraub and P. Price "Automatic text-independent pronunciation scoring of foreign language student speech" Proc. ICSLP 96, Philapdelphia 1996, pp.1457-1460.

[Rile98] M. Riley et. Al. "Stochastic Pronunciation modelling from hand-labelled phonetic corpora", pp109-116, Proc. ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Rolduc, May 1998.

[Schul98] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Large Vocabulary Speech Recognition", Proc. ICSLP98.

[STIL98] ESCA ETRW Workshop **STiLL** Speech Technology in Language Learning ,May 25-27 1998,Marholmen, Sweden

[Tranc99] I. Trancoso, C. Viano, I. Mascarenhas and C. Teixeira "On deriving rules for nativised pronunciation in navigation queries", EUROSPEECH 99

[VCom91]D. Van Compernolle, J. Smolders, P. Jaspers and T. Hellemans "Speaker Clustering for Dialectic Robustness in Speaker Indpendent Recognition", Eurospeech91, pp.723-726

[VODIS] VODIS, “Advanced Speech Technologies for Voice Operated Driver Information Systems”, EC Language Engineering Project LE 1-2277.

[Zava96] G. Zavagliakos, "Maximum A Posteriori Adaptation For Large Scale HMM Recognizers", Proc. ICASSP96, pp. 725-728

ACOUSTIC-PHONETIC MODELING OF NON-NATIVE SPEECH FOR LANGUAGE IDENTIFICATION

R. Wanneroy[1*], *E. Bilinski*[2], *C. Barras*[1], *M. Adda-Decker*[2], *E. Geoffrois*[1].

[1]DGA/CTA/GIP, 16 bis av. Prieur de la Côte d'Or, F-94114 Arcueil cedex
[2] LIMSI-CNRS, bat. 508, BP 133, F-91403 Orsay cedex

Abstract

The aim of this paper is to investigate to what extent non native speech may deteriorate language identification (LID) performances and to improve them using acoustic adaptation. Our reference LID system is based on a phonotactic approach. The system makes use of language-independent acoustic models and language-specific phone-based bigram language models. Experiments are conducted on the SQALE test database, which contains recordings from English, French and German native speakers, and on the MIST database, which contains non-native speech in the same languages uttered by Dutch speakers. Using 5 seconds of telephone quality speech, language identification error rate amounts to 10% for native speech and to 28% for non-native speech, thus yielding an important increase in error rate in the non-native case. We improve non-native language identification by an adaptation of the acoustic models to the non-native speech.

1 INTRODUCTION

In the field of automatic speech processing, intensive research activities have been devoted to speech recognition and transcription. With the growing interest in multilinguality and multilingual systems, language identification (LID) has become a research area of its own [5, 7]. In a multilingual context however speakers may use foreign languages for communication. Under such conditions, i.e. dealing with non-native speech input, system performances are known to decrease. Yet systematic evaluations of such degradation and research efforts to minimize them are still to be fostered.

Various information sources can be exploited in order to identify a given language: acoustic, phonemic, phonotactic, lexical, etc. In practice, for each information level specific resources and corpora are required for the languages to be modeled, and in most LID approaches only acoustic-phonetic and phonotactic models are used. The models are usually trained on native speech. Given the much greater spectral variability commonly observed in non-native speech, performance is expected to degrade when applying to such material.

Studying the impact of non-native speech on LID requires appropriate test material. Ideally a multilingual native speaker database and a multilingual non-native speaker database are required. Both corpora should be similar in style and recorded in comparable acoustic conditions. To our knowledge the MIST database is the first multi-lingual corpus gathering non-native speech; it contains recordings in English, French and German from Dutch speakers. Similar native speech material is provided by the multilingual corpora produced within the LE-SQALE project [6].

In the following, we describe the LID system used for the experiments. We present baseline LID results on native speech using the SQALE test database, and results on non-native speech using the MIST database; by means of these experiments we measure the impact of native versus non-native speech on LID error rates. Finally we investigate the effectiveness of acoustic model adaptation to handle non-native speech.

2 LID SYSTEM

The LID system used in the experiments is based on a phonotactic approach, with a single language-independent acoustic-phonetic decoder. This approach was chosen because, compared to language-specific acoustic modeling, it allows easier extension of the system to new languages, as there is no need

[*] financed by CS (Communications & Systems) company.

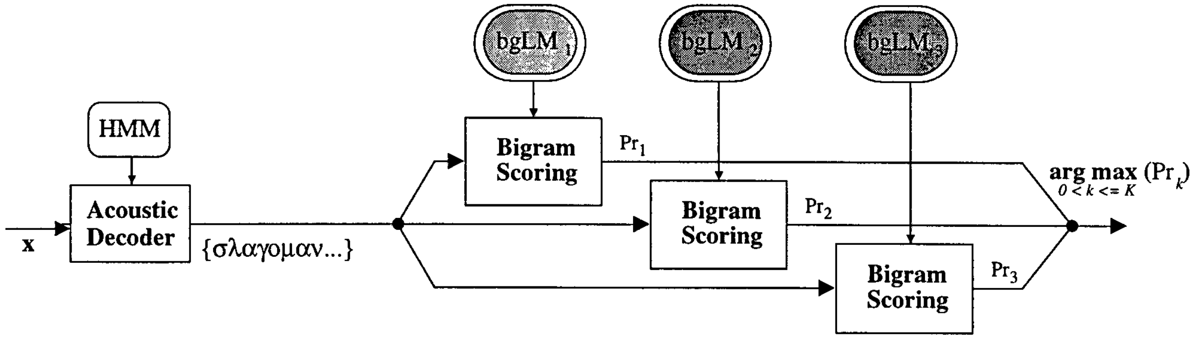


Figure 1: LID system using language-independent acoustic models and phone-based bigram language models

of specific phonetic knowledge of the new language or of a phonetic labelling of the training databases. The drawback is that it generally requires longer test segments to obtain optimal results as compared to acoustic-phonetic approaches. Previous work showed that the phonotactic approach LID results significantly improve when the test segment length goes from 10s to 45s [4].

The system is more extensively described in another article [1], where it is referenced as LI_HC (language-independent hierarchically clustered phone set). It is illustrated in Figure 1. It uses one single language-independent phone recognizer to label the speech input. The phone sequence output by this phone recognizer is then scored with language-dependent phonotactic models approximated by phone bigrams. The language providing the highest phonotactic probability is hypothesized.

2.1 Training database

The LID system was trained using the IDEAL corpus, which is a multi-language telephone speech corpus designed to support research on LID [3]. This corpus contains a large amount of speech (between 15 and 18 hours per language). The different languages were collected under the same conditions, and native speakers were recruited in their home countries. Data have been recorded for British English, Spanish, French and German. All speakers called the LIMSI data collection system ensuring the same recording conditions for the entire corpus. The IDEAL corpus contains about 300 calls for each language (i.e., international calls from native U.K., Spanish, and German speakers and national calls from native French speakers), 250 of them being used for acoustic and phonotactic model estimation (about 13 hours per language).

The calling script was designed to cover a variety of data types: 12 questions to elicit precise responses (7 general questions concerning the call and caller, and 5 prompts asking for times, dates, days of the week and months of the year), 18 items containing predefined texts to read, and 6 questions aimed at collecting spontaneous speech. The acoustic models were trained on all types of material, and the phonotactic models on the spontaneous speech part.

2.2 Front-end processing

The front-end processing consists in 12 MFCC plus the energy, augmented by their first and second order derivatives, i.e. a total of 39 coefficients every 10 ms. The same setting was used for processing test data, except that signal frequencies over 3.5 kHz were cut in order to be consistent with the training database which contains only narrow-band telephone speech.

2.3 Acoustic models

250 calls from IDEAL (about 9000 sentences, containing up to 13 hours of speech for each language) have been used for acoustic model training. First, 4 language-specific phone sets for English, French, German, and Spanish were trained. All acoustic models are three-state continuous density HMM of context-independent phones. Then a single multi-lingual set of 91 monophone models was obtained by an agglomerative hierarchical clustering of these 4 phone sets, using a measure of similarity between phones [1]. This phone set has proven to allow effective extension to new languages [4].

2.4 Phonotactic models

Phonotactic models were estimated on the spontaneous speech part of the 250 training calls which accounts for about 15% of the IDEAL corpus. For each language, an acoustic-phonetic decoding of the training database was performed using the multilingual phone set. The decoded phone strings are then used to estimate language-dependent bigram models for English, French and German.

3 TEST CORPORA

Experiments were conducted on the SQALE and MIST databases for LID results on native and non-native speech, respectively.

3.1 SQALE database

The development and test data of the SQALE project [6] were used for the native speech experiments. The 4-language (French, British and American English, German) speech database contains 400 sentences per language from 40 speakers, plus some diagnostic sentences which were not used in our experiments. Within the SQALE project the test sentences were chosen to give a reasonable spread of difficulty as determined by sentence length and perplexity. French, English (British or American) and German speakers were recorded reading newspaper texts from Le Monde, Wall Street Journal and Frankfurter Rundschau, respectively.

3.2 MIST database

The MIST database was developed by the TNO Human Factors Research Institute to support research in multi-linguality and non-native speech. 74 native Dutch speakers (52 male, 22 female) uttered 10 sentences in Dutch, and also for most of them in English, French and German: 5 sentences per language identical for all speakers and 5 unique sentences per language and per speaker. The text sources are the same as for the SQALE project concerning English, French and German. We used only unique sentences for evaluation on non native speech because identical sentences are not phonetically balanced over time. Finally, the selected part of the MIST database contains about 300 sentences per language.

4 EXPERIMENTAL RESULTS

We present LID error rates for each language as a function of sentence duration. Every second, the system takes a decision on the speech segment decoded so far. For a given test duration, only sentences longer than this duration were used. In order to reduce duration variability due to pauses and hesitations, the silences labelled by the recognizer are discounted from the sentence duration. For both test corpora, mean sentence duration is about 6 seconds. Few sentences are more than 8s long, and no significant LID results were obtained for segment durations over this duration.

4.1 Results on native speech

Identification results (on a second per second basis) on the SQALE database are provided in Figure 2. For 5 second segments, the global error rate amounts to 10%. This global rate does not show the disparity between languages; indeed, error rates of 16%, 3% and 10% are achieved for English, French and German speech, respectively. For all durations, results on French are significantly better than on the other languages. This might be attributed to the difference between French national and international telephone networks.

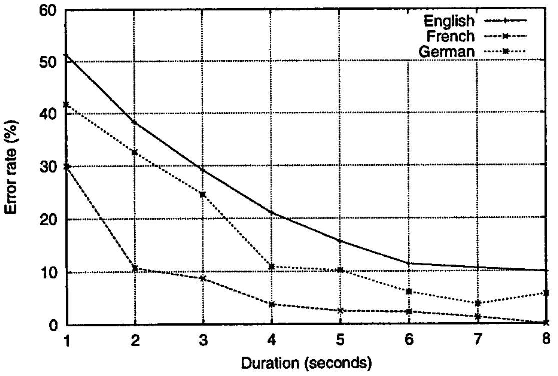


Figure 2: LID error rates for the native language task (SQALE database) as a function of segment duration.

4.2 Results on non-native speech

Similar experiments were conducted on non-native speech. The identification results using the three non-native MIST languages are illustrated in Figure 3. On 5 second segments, LID error rates for non-native English, French and German are 23%, 29% and 31%, respectively. The global LID error rate of the three non-native languages is 28%.

4.3 Comparison between native and non-native speech

The comparison of the identification results for native and non-native speech for each language is illustrated

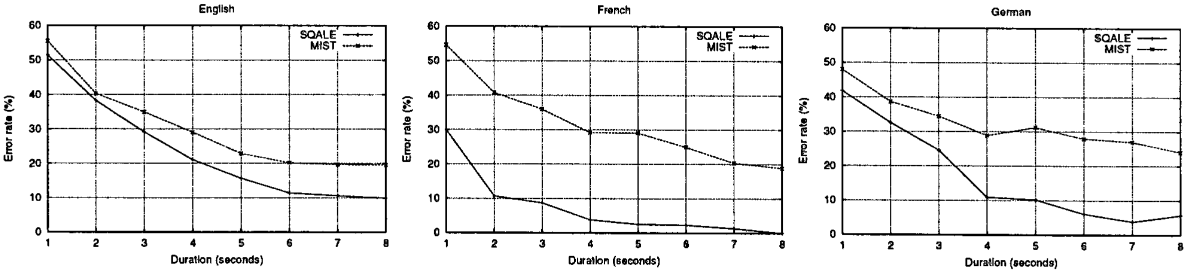


Figure 4: LID error rate comparison between native and non-native speech for English, French and German as a function of segment duration.

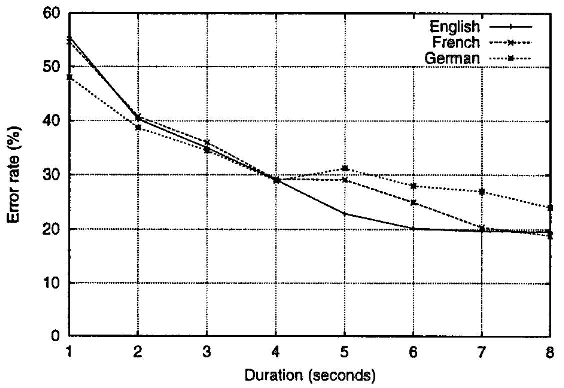


Figure 3: LID error rates for the non-native language task (MIST database) as a function of segment duration.

Table 1: Per language and global LID error rates on native speech (SQALE database) and non-native speech (MIST database) for 5 seconds of speech.

	SQALE	MIST	relative increase
English	16%	23%	×1.4
French	3%	29%	×10
German	10%	31%	×3.1
Global rate	**10%**	**28%**	**×2.8**

in Figure 4. For French and German, the non-native Dutch accent increases the error rates as expected. But error rate increase for non-native English, though significant, is much lower. The English phonotactic model seems to be more robust with respect to accent variation. Another more linguistically motivated conclusion consists in suggesting that Dutch speakers are best in speaking English as compared to French and German. For 5 second segments, the global error rate amounts to 10% for native speech and to 28% for non-native speech, showing an important increase in error rate (cf. Table 1).

4.4 Adaptation of acoustic models

Better results on non-native speech should be obtained after adaptating the LID system to the new conditions. Given the size of the available non native speech material (the MIST test database), an adaptation of the phonotactic models does not seem possible, and only acoustic models adaptation was tested. For a better use of the available data, the non-native MIST data were jack-knifed in 5 sets; the results were obtained by testing each set with acoustic models adapted on the remaining part of the database.

Each non-native sentence of the adaptation subset is aligned with the original prompt using the language-dependent acoustic models and produces a phone segmentation which is converted into the language-independent phone set. For each of the three non-native language, the acoustic models (including means, variances and weights of gaussians) are adapted towards the non-native acoustic realization of the phones. As a result, we get three sets of acoustic models. A weighting factor allows to control the degree of adaptation.

The LID system with adapted acoustic models is finally tested on the left-out fifth of the database. Each test sentence is decoded using the three adapted acoustic models in parallel with the original multilingual phone set, and the four phone sequences obtained are scored with the phonotactic models. The chosen language is the one with the highest global probability.

Figure 5 shows the global LID error rates after adaptation of the acoustic models. On 5 second segment, LID error rates of 21%, 22% and 27% are

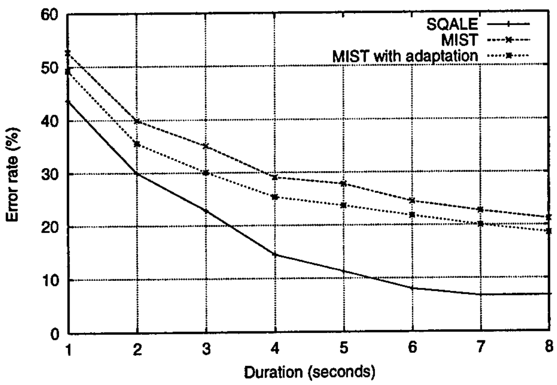


Figure 5: Gobal LID error rate for the native language task (SQALE database) and for the non-native language task (MIST database) before and after adaptation of acoustic models, as a function of segment duration.

achieved for non-native English, French and German respectively (these figures can be compared to those in Table 1). A 14% relative decrease of the global LID error rate is observerd for the three non-native languages (24% with adaptation vs. 28% without adaptation); despite the small size of the test set, this improvement can be shown to be significant using McNemar's test [2].

5 CONCLUSIONS

Experiments have been carried out with a phonotactic-based approach LID system on a 3-language task using native and non-native speech (SQALE, MIST corpora).

Using 5 seconds of telephone quality speech, LID error rate increased from 10% for native speech to 28% for non-native speech. Given the limited amount of test data, the test segment duration has been limited to a maximum length of 8 seconds, which stays far away from the typical durations (30s and more) for which the phonotactic LID approach performs best.

Adaptation of the acoustic model sets allowed to significantly reduce the error rate on non-native speech. Using the phonotactic approach, adaptation of the phonotactic models should be more efficient, but it could not be tested with the databases involved.

Needs for further investigation are obvious. Studying the effects of non-native speech on LID requires larger databases including more utterances of longer durations, more languages and various foreign accents. The development cost of such resources is of course a major issue. But the MIST database, even if only devoted to Dutch accent over a few European languages, was clearly an excellent starting point for the study of non-native speech, especially because of its matching with the already studied native SQALE database.

Acknowledgements

The authors wish to thank TNO for sharing the MIST database.

References

[1] C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel, "Language Identification with Language-Independent Acoustic Models", *Eurospeech'97*, pp. 55-58, Rhodes, Sept. 1997.

[2] L. Gillick, S.J. Cox, "Some statistical issues in the comparison of speech recognition algorithms", *Proc. ICASSP*, pp. 532-535, Glasgow, 1989.

[3] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J. Gangolf, J.L. Gauvain, "A Multilingual Corpus for Language Identification", *1st Int. Conf. on Language Resources and Evaluation*, pp. 1115-1122, Granada, May 1998.

[4] D. Matrouf, M. Adda-Decker, J.L. Gauvain, L. Lamel, "Comparing different model configurations for language identification using a phonotactic approach", *Eurospeech'99*, Budapest, Sept. 1999.

[5] Y.K. Muthusamy, E. Barnard, R.A. Cole, "Reviewing Automatic Language Identification", *IEEE Signal Processing Magazine*, Oct. 1994.

[6] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland, "Multilingual large vocabulary speech recognition: the European SQALE project", *Computer Speech and Language*, **11**, pp. 73-89, 1997.

[7] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. on SAP*, 4(1), pp. 31-34, Jan. 1996.

How foreign are "foreign" speech sounds? Implications for speech recognition and speech synthesis

Anders Lindström & Robert Eklund*

{Anders.P.Lindstrom,Robert.H.Eklund}@telia.se

*Telia Research AB, Room B324, S-12386 Farsta, Sweden

ABSTRACT

This paper reports results from a production study which shows in what ways the traditional Swedish phone set is expanded with phones similar to or approximating phones from other languages than Swedish in everyday speech. The inclusion of such sounds – here called *xenophones* – has implications for both automatic speech recognition and speech synthesis systems, especially in polylingual environments, which are discussed in the paper.

1. INTRODUCTION

In speech technology systems there is an increasing interest in issues such as dialectal variation, cross-language applications, handling of foreign accents et cetera. This problem is becoming more acute in an increasingly internationalized world, where people tend to speak more than one language, and also tend to ask for services that pay little or no attention to national or language borders.

A hitherto somewhat neglected problem that constitutes an important issue in the development of such multilingual applications is dealing with the fully normal inclusion of "foreign" speech sounds in the pronunciation of foreign names and words. Such speech sounds can be said to expand the phone inventory of the (native) language in question, a phenomenon observed in at least some languages, such as Swedish [5,6,7,10,11]. An example from Swedish would be the voiceless dental fricative [θ] (the first sound in the name "Thatcher"), which is not considered part of the Swedish phonemic inventory, but is nevertheless produced by approximately 50 percent of the population when pronouncing English words or names containing this sound in otherwise Swedish sentence contexts [6,10,11].

With a growing awareness of the need for multilingual automatic services (cf. e.g. [3]), the handling of language users' less constrained pronunciation becomes something of a *sine qua non*.

1.1. The Xenophone Problem

As was mentioned above, it has been shown that Swedes' pronunciation of names or words of foreign origin often exhibit sounds that are not part of what is considered the Swedish phoneme inventory. Such "added" sounds do not have a phonemic function in Swedish, and must therefore be attributed a particular status in the system. Even though they are not phonemes – or allophones of Swedish phonemes – they are clearly part of the *phone* sets of individual Swedish speakers. Hence, we suggested the term *xenophones* [6], i.e. "foreign phones", to denote such sounds.

Appropriate treatment of this phenomenon is likely to influence the performance of any speech recognition or synthesis system. For both these types of applications, expansions of the phone set are required. What is also apparent in the results reported in Eklund & Lindström [ibid.], is that the nature of this xenophonic expansion depends on the particular sound in question (among other things). This leads into the field of phonological acquisition, and more specifically, into the field of second language acquisition (SLA) research. The phonological processes involved when approaching a foreign language have been discussed in detail since long (e.g. [8,9]), and SLA research definitely provides valuable insight with regard to what factors might be at play. However, we would like to argue that although the phonological foundation is the same in xenophonic expansion and SLA, xenophones present a different problem since we are facing a different situation. Within SLA, the goal of the subject(s) is to master an entire target language, often in a target language context, whereas in the case of xenophonic expansion, the subjects simply include words of foreign origin in native-language sentences, mostly within fully native-language contexts. Thus, the entire communicative goal may be considered different, and this in turn should affect the actual rendering of the linguistic items in question. Indeed, as we shall see, this is supported by a some of our observations.

As discussed in Eklund & Lindström [7,10], a number of underlying factors can be assumed to be involved in governing the degree of adjustment. See Figure 1.

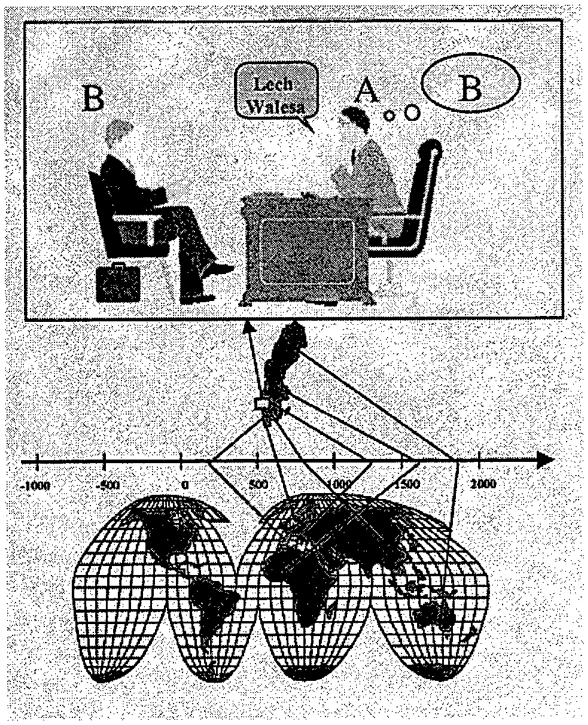


Figure 1: The language user in a typical situation. When speaker A is pronouncing a name in speaker B's presence, a number of factors are affecting the phonetic rendering of the name, such as the name's country of origin, the time it was introduced in speaker A's community, what channel it passed through, A's knowledge of B's language competence and other factors.

These include – but are not limited to – the speaker's competence and performance capabilities with respect to the source language, the speaker's expectations of the listener's competence, the relative social status of speaker and listener, the socio-cultural distance to the country of origin, recency and frequency of the lexical item in question, and similarities and dissimilarities between the phonological systems involved.

2.2. Previous Work

Despite the fact that the problem is crucial, or even central in some languages, and despite the fact that references are found that date back to the 16th century, very little actual work on the phenomenon has been reported.

Maddieson [12] briefly discusses the phenomenon but simply refers to the phones in question (in several languages) as "anomalous" segments.

Abelin [1] discusses how to represent pronunciation of foreign (mainly English) words in *Svensk Ordbok.* She concludes that the English diphthongs [eɪ] and [oɪ] can be approximated with the Swedish sequences [ej] and [oj], respectively, but that the English diphthongs [əʊ] and [aʊ] are harder to accommodate. The English phone [z] is more or less always pronounced as [s] in Swedish, and the English alveolars [r, t, d, n] are normally realized as dentals in Swedish.

Eklund & Lindström [6] describe what English phones Swedes actually use in their speech, and show that a large proportion of Swedish speakers include "non-Swedish" sounds in their production system when pronouncing English words and names. Eklund & Lindström also describe the inclusion of xenophones into the Telia Research concatenative synthesizer.

Möbius et al. [13] mention that the German version of the Bell Labs multilingual TTS system has been augmented with phonetic units outside the German phone inventory in order to cover English and French speech sounds.

2. METHOD

In order to acquire information and knowledge concerning Swedish speakers' usage of xenophones, and also, to some extent, insight in their expectations on xenophone usage, a production study was conducted. The rationale for looking at production data, we argue, is that knowledge may be gained in several dimensions: Which English phones have an effect of the Swedish subjects' productions? What is the nature of this effect—is the phone repertoire extended or does some kind of segmental mapping take place? Even if a speaker does not produce an English name or word in an accent-free manner, he or she might still do something that clearly lies outside the Swedish phone inventory. By producing something that is neither Swedish nor English, as it were, the speaker is indicating an awareness of the difference between the English pronunciation and a fully rephonematized pronunciation (i.e., "translating" the English sound into its phonetic "counterpart" in Swedish). This provides important information in the "attitude dimension", insofar as it shows that even speakers who do not fully master the production of English sounds might expect these sounds to occur in particular words.

2.1. The Linguistic Material

A set of twelve sentences was constructed containing the 15 English speech sounds [tʃ, dʒ, ʃ, ʒ, θ, ð, z, ɚ, ɫ, w, aɪ, eɪ, əʊ, juː, æ]. The two non-English (and non-Swedish) sounds [x,aː] were also included in the material. All these sounds were chosen so that they would differ phonetically from Swedish speech sounds to varying degrees, and so that none of them would be included in any traditional description of the Swedish phonological system.

The phones were included in commonly known names and words in twelve fully natural Swedish sentences and it was assumed that the words and names in which the xenophones appeared would be known by the bulk of the subjects.

Two example sentences from the material are given below.

Många har Roger Moore som favorit i rollen som James Bond.
("A lot of people prefer Roger Moore's interpretation of James Bond")

Intercity-tåget gick direkt från Aachen till Baden-Baden.
("The Intercity train went straight from Aachen to Baden-Baden")

2.2. Recordings and Subjects

The sentences were included in a much larger session of linguistic material recorded to train the Telia/SRI Swedish speech recognizer as a part of the *Spoken Language Translator* (SLT) project [2,14]. The material was presented under the heading 'Kändisar' (Celebrities), and it can be assumed that subjects were unaware of the fact that their pronunciation was the object of study.

The subjects were all Telia employees or relatives of Telia employees. The age span was 15 to 75. Hi-fi recordings were obtained of more than 460 subjects on 40 different locations covering the whole of Sweden, so that data from all major dialect areas were obtained. In this way a total of approximately 29,000 xenophone tokens were collected. The subjects also filled in forms, providing information concerning educational level, regional origin and so on.

2.3. Evaluation

Three phonetically trained native speakers of Swedish, with an above-average knowledge of English, transcribed the target phones, using a fairly narrow allophonic transcription scheme. So far, 15,202 potential xenophone tokens have been evaluated.

3. RESULTS

Figure 2 shows the proportional distribution of the subjects' productions of the speech sounds [aː, aɪ, eɪ, əʊ, juː, æ, tʃ, dʒ, x, ʃ, ʒ, θ, ð, z, ɚ, ɫ, w], where each instance of these has been assigned to one of three categories along the awareness and fidelity dimensions.

Category 1 corresponds to high awareness coupled with high fidelity, production-wise.

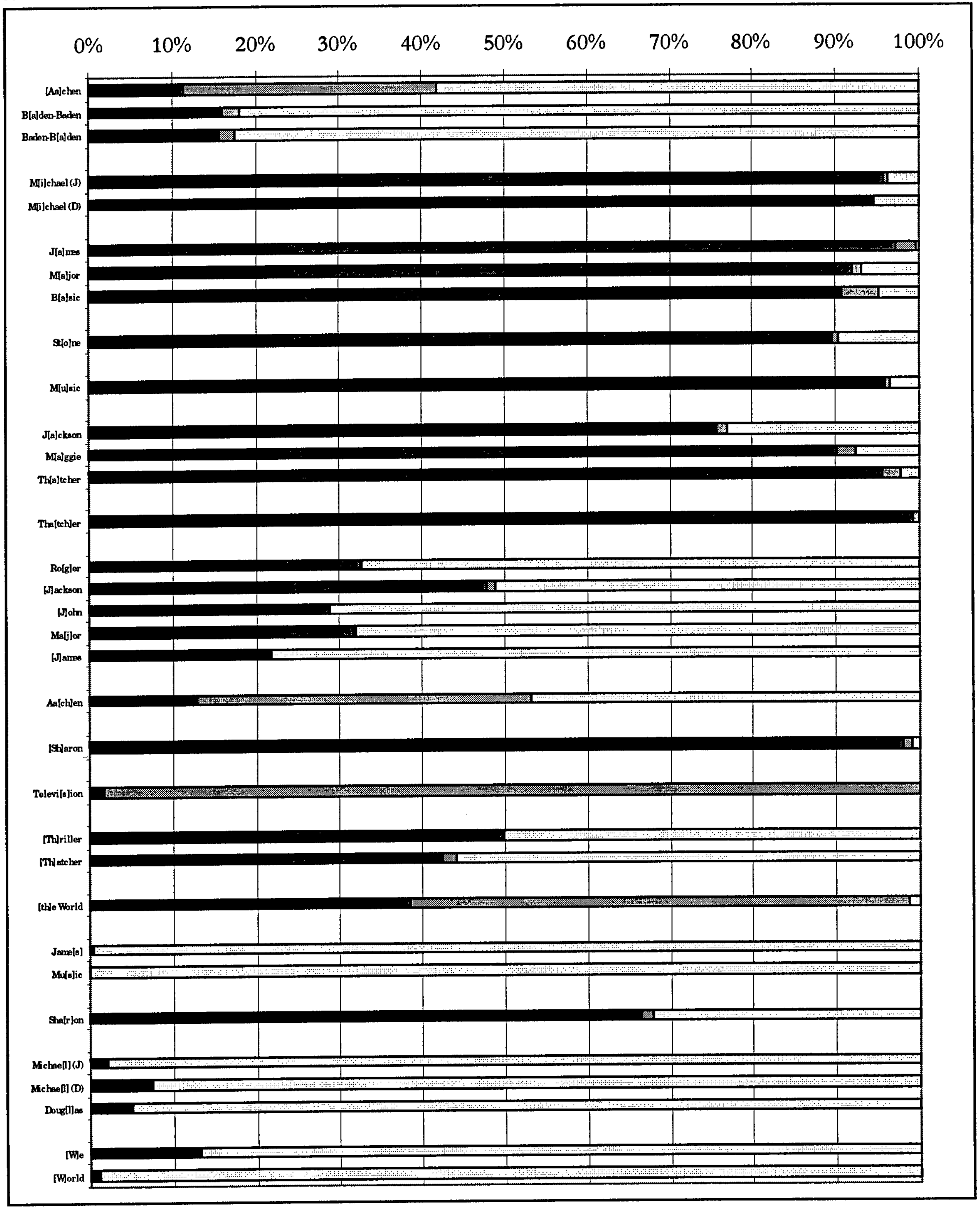
Category 3 indicates low fidelity, and probably low awareness, although it may also be the case that some speakers deliberately rephonematize (for normative reasons).

Category 2, high awareness and low fidelity, is interesting, since it represents those speakers who are apparently aware that *something* foreign should be going on, but fail to produce a good enough approximation of the "target" speech sound. Speakers in this category can certainly cause considerable problems for ASR systems.

As can be seen in Figure 2, the distribution over the three categories differs considerably as a function of target phone, and even as a function of each individual "lexical item". It is interesting to note that voiced fricatives are more or less non-existing, despite the fact that are easy to produce, whereas the more "remote" phones (from a number-of-phonetic-features perspective), from a Swedish point of view, e.g. dental fricatives, are produced by a large number of subjects.

A subset of the data presented here has also been evaluated with respect to which underlying factors might explain the differences in use of xenophones.

Figure 2: For each target English speech sound and each occurrence in the read sentences, the proportional distribution of the Swedish subjects' productions is shown. Based on the similarity between the produced sound and the target phone, the different productions are assigned to one of three categories along two dimensions, the *awareness dimension* (to what extent people are aware of the difference between Swedish and English pronunciation), and the *fidelity dimension* (how well they succeed in the production of the foreign sounds). The first category (magenta/dark grey) corresponds to a high awareness among the subjects coupled with a high capability in rendering a sound close to the one in the source language. The second category (green/middle grey) corresponds to the case where the subjects were apparently aware that something "non-Swedish" would be appropriate, but failed to produce a good approximation. The third category (yellow/light grey) corresponds to full adjustment to Swedish.



Lindström and Eklund [11] showed that age seems to be one factor that systematically affects the productions, in such a way that the youngest and oldest subjects generally produce relatively more category 2 and 3 productions than do the other subjects.

In the same study, no significant gender differences were found, nor were there any systematic regional differences. The last result, however, may be due to lack of control for the variable "educational level", and re-evaluation of the data with that in mind is in the works.

4. DISCUSSION

Xenophones can be discussed and studied from several different angles. From a theoretical perspective, the underlying theoretical issues xenophones raise mainly concern general phonological acquisition, relating to, without being similar to, SLA research.

As indicated in Figure 1, there are a number of underlying factors that can be assumed to be at play in determining the choice of the speaker's pronunciation strategy, and we believe that we have shed some light on the issue of what speakers do when solving this task of finding the socially acceptable level along the awareness/fidelity dimension.

From a theoretical side, the "foreignness" of such sounds can be discussed. If most Swedes use certain sounds in everyday conversation, and/or expect them to be used, how "foreign" are they in the language community? Moreover, in a world that is characterized by increasing international communication – economical, cultural, social – such cross-breeding between languages can be expected to become more and more frequent.

From a more practical side, there are a number of consequences that these observations are bound to have for automatic speech recognition and speech synthesis.

4.1. Implications For Recognition

A recognizer is facing the entire variety of speech sounds within a given speech community, and the modelling of what it can be expected to hear boils down to a few crucial issues.

First, the standard view on what the Swedish phone set looks like must be reconsidered, since it obviously to a large degree contains sounds normally not considered "Swedish", despite the fact that a large number of Swedish speakers do use them in normal conversation.

To complicate matters further, a word/name of foreign origin and containing foreign – or foreign-similar – sounds can appear in an otherwise Swedish sentence, which means that the recognizer needs to handle phones from (at least) two languages at once. Within the SLT project, a recognizer that is able to handle English and Swedish was developed [4,15,16]. The recognizer is capable of recognizing the odd Swedish word inside an otherwise English sentence, and vice versa.

Another issue is exactly how acute a problem xenophones present to a recognizer. This, of course, depends heavily on the context and discourse. An application like automatic handling of film ticket purchasing would surely need to cope with a large number of xenophones, since most English film titles are not translated into Swedish. Within other domains, such as bookings of summer houses in the Stockholm archipelago, xenophones are not likely to occur at all. Thus, xenophone inclusion for a given application is also an empirical issue.

4.2. Implications For Synthesis

As opposed to recognition, where the entire variety needs be considered and catered for, a synthesizer probably only needs to cover *one* acceptable variety. The operative word here, of course, is "acceptable". Although it is our belief that a production study provides information in the acceptability domain insofar as it can be assumed that users of speech synthesis systems will be less prone to accept a synthesizer with a lower level of competence than themselves, the only safe method to gain insight in the acceptability domain would be to conduct a perception study. One such method would be to play back to subjects the obtained recordings and ask them rank the pronunciations along a few dimensions, such as intelligibility, "intelligence", pleasantness and so on. It is our belief that a low inclusion level of xenophones might not primarily show up in the intelligibility dimension, but rather present itself to listeners as a synthesizer with a low educational level.

Another problem to consider is that "maximizing" in the xenophone dimension might leave certain listeners behind, especially concerning languages that are not so commonly known as English (e.g. French, German or Russian) and that an appropriate level must be found. It can be assumed that choosing too "high" a level will signal an attitude which would be perceived as high-browed and obnoxious. This, too, needs more studies.

To the best of our knowledge, few attempts to include xenophones in synthesizers have so far been made. As mentioned above, Eklund & Lindström [6] report the inclusion of English xenophones in the Telia Research research synthesizer and Möbius et al. [13] mention the inclusion of a few English and French sounds in the German version of the Bell Labs multilingual TTS system.

4.3. Future Research

Apart from the perception studies mentioned above, a deeper look into the phonological-regional dimension is needed. The rationale for doing this is that one thing one would want from an intelligent recognizer is that it possess a certain level of predictive power, so that it could "tune in" to a particular speaker's use of xenophones (and idiosyncratic speech behavior in general). However, our observations so far do not provide much hope in that dimension, since the speakers generally do not exhibit a high degree of consistency in their use of xenophones. For example, a phrase like *Diana and Charles* (from the material) may be pronounced with xenophones on *Diana* but not on *Charles*, or vice versa. Thus, our studies so far indicate that xenophone inclusion may appear spot-wise, rather than consistently. However, this asks for more research.

Another thing that awaits studies is to what extent prosodic signaling is employed. Some subjects signaled awareness of the foreignness of the names and words by using a prosodic realization that is influenced by the source-language, in this case English, either in addition to, or independently of, the use of xenophones. So far, we have not conducted any formal studies of this phenomenon, and the benefits from such knowledge of course require that recognizers make use of prosody, something which currently is not done, at least not to any larger degree.

Another factor to be studied further is the role of orthography, something we have tried to normalize for by including the same sounds with different spelling (i.e., the voiced affricate [dʒ] was presented both in the name *James* and in the name *Roger*). It proved to have some effect [6], but more data are needed before any far-reaching conclusions made be drawn concerning the role of orthography.

Finally, an obvious factor to study is the speakers' educational level. It goes without saying that previous and close familiarity of foreign languages affect the pronunciation, as well as one's expectations on how names and words of foreign origin "should" be pronounced. Such studies are underway, and will be reported in future work.

5. ACKNOWLEDGEMENTS

The authors thank Per Sautermeister for help with the listening task. Also thanks to Malin Ericson, who made comments on draft versions of the paper.

6. REFERENCES

1. Abelin, Å. 1985. Om uttalsmarkering och uttalsregler i svensk ordbok. *Rapporter från Språkdata (21)*, Gothenburg University, Department of Computational Linguistics, Gothenburg.
2. Becket, R., P. Boullion, H. Bratt, I. Bretan, D. Carter, V. Digalakis, R. Eklund, H. Franco, J. Kaja, M. Keegan, I. Lewin, B. Lyberg, D. Milward, L. Neumeyer, P. Price, M. Rayner, P. Sautermeister, F. Weng & M. Wirén. 1997. *Spoken Language Translator: Phase Two Report.* Telia Research AB and SRI International.
3. Billi, R. n.d. Interview published in *Le Journal – The Journal of Record for Human Language Technology.* http://www.linglink.lu/lejournal/article.asp?articleIndex=628
4. Digalakis V. & L. Neumeyer. In press. Multiple Dialects and Languages. In M. Rayner, D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). *The Spoken Language Translator.* Ch. 18, pp. 307–318. Cambridge University Press.
5. Eklund, R., J. Kaja, L. Neumeyer, F. Weng & V. Digalakis. In press. Porting a Recognizer to a New Language. In M. Rayner, D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). *The Spoken Language Translator.* Ch. 17, pp. 297–306. Cambridge University Press.
6. Eklund, R. & A. Lindström. 1998. How To Handle "Foreign" Sounds in Swedish Text-to-Speech Conversion: Approaching the 'Xenophone' Problem. *Proc. of ICSLP 98*, Sydney, November 30–December 5. Paper 514, Vol. 7, pp. 2831–2834. CD-ROM available from Causal Productions Pty Ltd, PO Box 100, info@causal.on.net.
7. Eklund, R. & A. Lindström. 1996. Pronunciation in an internationalized society: A multi-dimensional problem considered. *FONETIK 96, Swedish Phonetics Conference, Nässlingen, 29–31 May, 1996. TMH-QPSR 2/1996*, 123–126.
8. Flege, J.E. 1987. Effects of Equivalence Classification on the Production of Foreign Language Speech Sounds. In James, A. & J. Leather (eds.). *Sound Patterns in Second Language Acquisition*, Foris Publications.
9. Hammarberg, B. 1990. Conditions on Transfer in Second Language Phonology Acquisition. In Leather, J. & A. James (eds.), *New Sounds 90, Proc. of the 1990 Amsterdam Symposium on the Acquisition of Second-Language Speech.* University of Amsterdam.
10. Lindström A. & R. Eklund. 1999. Xenophones Revisited: Linguistic and other underlying factors affecting the pronunciation of foreign items in Swedish. *Proc. of ICPhS 99*, San Francisco, August 1–7. Paper 0708.
11. Lindström A. & R. Eklund. 1999. [jɑ̀ːmes] or [dʒɛɪmz] or Perhaps Something In-between? Recapping Three Years of Xenophone Studies. Gothenburg Papers in Theoretical Linguistics, 81. *Proc. Fonetik 99*, The Swedish Phonetics Conference, June 2–4 1999, pp. 109–112.
12. Maddieson, I. 1984. *Patterns of sounds*, Cambridge University Press.
13. Möbius, B., R. Sproat, J.P.H. van Santen & J.P. Olive. 1997. The Bell Labs German Text-to-Speech System: An Overview. *In Proc. ESCA. Eurospeech 97, Rhodes, Greece*, ISSN 1018–4071, pp. 2443–2446.
14. Rayner, M., D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). In press. *The Spoken Language Translator.* Cambridge University Press.
15. Weng, F. In press. Language Modeling for Multilingual Speech Translation. In M. Rayner, D. Carter, P. Bouillon, V. Digalakis & M. Wirén (eds.). *The Spoken Language Translator.* Ch. 16, pp. 281–296. Cambridge University Press.
16. Weng, F., H. Bratt, L. Neumeyer & A. Stolcke. 1997. A Study of Multilingual Speech Recognition. *Proc. Eurospeech*, pp. 359-362, Vol. 1, Rhodes, Greece.

CLUSTERING OF CONTEXT DEPENDENT SPEECH UNITS FOR MULTILINGUAL SPEECH RECOGNITION

Bojan Imperl
University of Maribor
Smetanova 17, 2000 Maribor, SLOVENIA
E-mail: bojan.imperl@uni-mb.si

ABSTRACT

The paper addresses the problem of designing a language independent phonetic inventory for the speech recognisers with multilingual vocabulary. A new clustering algorithm for the definition of multilingual set of triphones is proposed. The clustering algorithm bases on a definition of a distance measure for triphones defined as a weighted sum of explicit estimates of the context similarity on a monophone level. The monophone similarity estimation method based on the algorithm of Houtgast. The clustering algorithm is integrated in a multilingual speech recognition system based on HTK V2.1.1. The experiments were based on the SpeechDat II databases[1]. So far, experiments included the Slovenian, Spanish and German 1000 FDB SpeechDat (II) databases. Experiments have shown that the use of clustering algorithm results in a significant reduction of the number of triphones with minor degradation of word accuracy.

1. INTRODUCTION

The development of speech technology in the last few years raised an interest in the research of the multilingual speech recognition. In order to reduce the complexity of a multilingual recogniser and to reduce the cost of a cross-language transfer of speech technology, the development of methods for the definition of the multilingual phonetic inventories is of increasing concern.

The definition of the multilingual phonetic inventories by exploiting similarities among sounds of different languages is a promising approach. First attempt was reported in [1]. Here the multilingual phonetic inventory, consisting of language-dependent and language-independent speech units, was defined using the data-driven clustering technique. Other attempts based on different distance measures and clustering techniques also followed [2,3,4,5], however, all the work so far was focused on the context independent phoneme modelling (monophones). These experiments have shown that the transition from language dependent monophone set to multilingual inventory of monophones may result in a degradation of recognition accuracy due to the lack of acoustic resolution of the multilingual phoneme set.

The transition from the context independent to context dependent phoneme modelling seems inevitable in order to improve the performance of multilingual speech recognition systems, i.e. the speech recognisers with multilingual vocabulary. The development of a method for the definition of the multilingual set of context dependent phoneme models requires the definition of new clustering criteria.

In this paper, a clustering algorithm for the definition of multilingual set of context dependent phoneme models (triphones) is proposed. The clustering algorithm bases on a distance measure for triphones defined as the combination of explicit estimation of the similarity of the phonemes of left and right contexts and the central phonemes.

2. TRIPHONE DISTANCE MEASURE

The crucial problem concerning the use of triphone modelling is large number of triphone models, which requires large amounts of training data. Since the amount of training data is usually limited many of the triphone speech units are rarely or even never seen during the training. For this reason the direct implementation of the distance measures that were defined for the

[1] The use of SpeechDat databses was enabled by the Siemens AG and the Universitat Politecnica de Catalunya.

monophones, such as [1, 2, 3, 4] is not appropriate for the definition of multilingual set of triphones.

Our definition of the distance measure for triphones bases on the fact that the triphone is "a monophone in a certain context". Therefore, the similarity of two triphones can be estimated also indirectly - by explicitly estimating the similarity of both central phonemes, both left-context phonemes and both right-context phonemes. The similarity of two triphones l_1-c_1+r_1 and l_2-c_2+r_2 (*l*, *c* and *r* denote the left context - phoneme, right context - phoneme and the central phoneme, respectively) was therefore defined as:

$$S(l_1\text{-}c_1+r_1,\ l_2\text{-}c_2+r_2) = L\,s(l_1,l_2) + C\,s(c_1,c_2) + R\,s(r_1,r_2) \qquad (1)$$

where *s* denotes the similarity of two phonemes, *L, C, R* are the weights for setting the influence of each phoneme - level similarity estimates, and $S(l_1\text{-}c_1+r_1,\ l_2\text{-}c_2+r_2)$ is the resulting similarity of both triphones.

Such definition of distance measure for triphones can be based on any type of phoneme-distance measure (*s* in Equation 1). In our case, the phone-distance measure was defined as suggested in [1]:

$$s(f_i,f_j)=s(f_j,f_i)=$$
$$\frac{1}{2}\sum_{k=1}^{N}\left[c(f_i,f_k)+c(f_j,f_k)-\left|c(f_i,f_k)-c(f_j,f_k)\right|\right]$$
$$1\leq i,j\leq N,\quad i\neq j \qquad (2)$$

where $s(f_i,f_j)$ denotes the similarity between phonemes f_i and f_j, *N* is the number of phonemes, $c(f_i,f_k)$ is the number of confusions between phonemes *i* and phone *j*.

Described definition of distance measure for triphones has two major advantages. First it offers an accurate estimation of a triphone similarity (similarity of triphones is likely to be higher in a matching context and vice-versa). Next, such definition can provide a reliable estimation of similarity between triphones even in case of "rare" or "unseen" triphones.

3. CLUSTERING ALGORITHM

Having defined the distance measure for the triphones, the clustering algorithm for automatic identification of the triphones that are similar enough to be equated across the languages was defined.

A group of triphones is equated if an average distance among all triphones from the group is less than a predefined threshold *T*. Average distance among *M* triphones was defined as:

$$S(\varphi_1,\varphi_2,\mathrm{K}\,,\varphi_M)=\frac{\sum_{k=1}^{M}\sum_{l=1}^{M}S(\varphi_k,\varphi_l)}{\sum_{k=1}^{M}k}$$
$$\varphi_k,\varphi_l\in(\varphi_1,\ \varphi_2,\ \ldots,\ \varphi_M),\ k\neq 1 \qquad (3)$$

where φ_k denotes the triphone l_k-c_k+r_k, (φ_1, φ_2, … , φ_M is the group of triphones, S (φ_1, φ_2, … , φ_M) is the average distance among all triphones from the group (φ_1, φ_2, … , φ_M). To find all groups of triphones that complies with the condition from the Equation (3), the following 2-stage search algorithm was applied.

In the first stage, a list of most similar phonemes (poly-phonemes) was defined using the method described in [1]. A partial list of poly-phonemes covering all three languages is given in Table 1.

n	Slovene	German	Spanish
1	**a**	**a**	**a**
2	**O**	**O**	**o**
3	**n**	**n**	**n**
4	**l**	**l**	**l**
5	**t**	**t**	**t**
6	**m**	**m**	**m**

Table 1. A partial list of poly-phonemes for the Slovene, German and Spanish language.

In the second stage, the groups of triphones to be equated were identified. The search for these groups was limited to the classes of triphones consisting of triphones with the phonemes of the same poly-phoneme as the central phoneme. For example, the search for the similar triphones was first started among the triphones of all three languages with either Slovenian phoneme **a**, German phoneme **a** or Spanish phoneme **a** as the central phoneme. Next, the search for the groups of similar triphones continued among the triphones with either Slovenian phoneme **O**, German phoneme **O** or Spanish phoneme **o** as the central phoneme, etc. Such limitation of search has proven to significantly improve the convergence of the algorithm for the identification of the groups of similar triphones due to the large number of triphones

This clustering algorithm outputs the list of triphones that are similar enough to be equated across the languages. The unlisted triphones remain language specific. The degree of equated triphones can be adjusted by the threshold *T*. The value of *T* was derived experimentally (values are given with the experimental results).

4. BASELINE RECOGNISER

The speech recognition system was based on HTK V2.1.1 with modified frontend module for enhancing the speech recognition robustness. The acoustic feature vector produced by the frontend module consisted of 24 mel-scaled cepstral, 12 Δ - cepstral, 12 ΔΔ - cepstral, high pass filtered energy, Δ - energy and ΔΔ - energy coefficients. This feature vector was processed using the algorithms for maximum likelihood channel adaptation [8] and linear discriminant analysis [8].

Such frontend module was chosen due to the results of previous tests on connected digits recognition task with 99 speakers of the Slovene speech database SNABI and tests on isolated digits recognition task with the databases SNABI and Voice-Mail (German).

The baseline speech recognition system consisted of three language specific recognisers (Slovene, German and Spanish) operating in parallel. The 3-state left-right topology was selected. The triphone models were initially built with 1 Gaussian mixture component per state. All together 24173 triphone models were defined (Sl.-7146, Ge.-12279,Sp.-4748). Parameter tying using the tree-based clustering algorithm (as implemented in the HTK) reduced the number of triphone models to 13074 (Sl.-3517, Ge.-6517,Sp.-3040). At the end the number of Gaussian mixture components per state was augmented to 32.

In the multilingual experiments, the three language specific recognisers operated in parallel using either three language specific model sets or one multilingual set of triphones where many of language specific triphones are tied and used by all three recognisers.

5. SPEECH DATABASES

The experiments were carried out using the speech databases produced in the framework of the SpeechDat II project [7]. These databases provide a realistic basis for developing voice driven teleservices and multilingual systems. The following SpeechDat databases were used:

- ♦ Slovenian 1000 FDB SpeechDat(II) [6],
- ♦ German 1000 FDB SpeechDat(II),
- ♦ Spanish 1000 FDB SpeechDat(II).

In all cases, the corpuses contained utterances of 1000 speakers. 800 speakers were used for the training and the remaining 200 speakers were used for the testing of the system. In all experiments the train and test sets were defined as recommended in SpeechDat II project specification.

Only 80 - 95 % of all utterances were useful for the experiments. Remaining utterances were skipped due to the following reasons:

- – unusual pronunciation of digits,
- – incomplete utterances (speech was cut off at the beginning or end of the utterance),
- – unexpected utterances (background noise, comments, ...).

The system was trained using all corpuses of the train set, while for the testing the corpuses W1-W4 of all three databases, containing phonetically reach words, were used (total of 2252 utterances containing 1960 different words).

6. EXPERIMENTAL RESULTS

The baseline recogniser was tested in monolingual and in multilingual mode of operation, where the three language specific recognisers operated in parallel.

The recogniser performance for the monolingual tests is given in the Table 2.a. The word accuracy (*WA*) is listed for each language. The performance of the recogniser using the triphone models with 1 Gaussian mixture component per state (models: tri1) was low. Augmenting the number of Gaussian mixture components to 32 (models: tri32) significantly improved the word accuracy. The transition from the monolingusl to the multilingual mode of operation (Table 2.b) did not significantly degrade the recognizer performance. In most cases the recognizer correctly recognized the language. Errors in language identification usually ocured for the words that were already misrecognized in the monolingual tests. Therefore the errors in language identification did not cause additional errors in word recognition. The language identification rate (*LI*) was high for both types of triphone models and the word accuracy of multilingual tests approximately equals to the average word accuracy of the monolingual tests.

a)

models	WA		
	SL	ES	DE
tri1	67.51%	78.58%	76.77%
tri32	88.25%	93.91%	92.51%

b)

models	WA	LI
tri1	71.99%	91.61%
tri32	91.52%	93.10%

Table 2. The baseline recogniser performance for the monolingual tests (a) and for the multilingual tests (b).

Experiments with multilingual set of triphones were carried out for the recogniser with 13074 models and 1 Gaussian mixture component per state. The word accuracy was therefore much lower than it would have been in the case of models with 32 Gaussian mixture components per state. However, the purpose of the experiments was to determine the optimal values of the clustering parameters (weights *L,C,R* and threshold *T*) and to compare the performance of the multilingual triphone set to the performance of monolingual triphone sets running in parallel. Augmenting the number of Gaussian mixture components per state from 1 to 32 would improve the performance of the multilingual triphone set in the similar way as it did for the monolingual triphone set (Table 2).

The clustering algorithm was started at different values of weights *L,C* and *R* (see Equation 1) and at different threshold values (*T*) producing the multilingual triphone sets of different sizes. The performance of the recogniser using various multilingual triphone sets is given in the Tables 3.a, 3.b and 3.c.

Beside the word accuracy and the language identification rate, the global compression rate [4] was also followed. The global compression rate (*GCR*) was defined as:

$$GCR = \sum_{i=1}^{N} c_i \frac{M_i}{T_i} \tag{4}$$

where L is the number of languages, T_i is the number of trainable models in language i, M_i is the number of merged models in language i and c_i is the ratio between the number of trainable models in language i and the number of trainable models in L languages.

The weights *L,C* and *R* were first set to the the values *L=1,C=0,R=1* (Table 3.a) . This way the similarity of both central phonemes did not have any influence to the resulting similarity of both triphones. The search for the groups of similar triphones was limited to the classes of triphones consisting of triphones with the phonemes of the same poly-phoneme as the central phoneme. Therefore the similarity of both central phonemes has already been considered during the search for the groups of similar triphones.

The use of multilingual set of triphones (models: tri1C) produced at weights *L=1,C=0,R=1* can reduce the total number of triphones of the baseline system (models: tri1), but it also results in a decrease of word accuracy and language identification rates in case of multilingual experiments (results from Tables 3 (WA-MULTI) are also shown on Figure 1). In best case the *GCR* of 24.19% is achieved at approximately 1% decrease of *WA* rate and more than 5% decrease of *LI* rate. Using the multilingual set of triphones for the monolingual experiments have shown an improvement of the word accuracy in case of Slovene language for the threshold values larger than 100 (results from Table 3.a (WA-SL) are shown also on Figure 1).

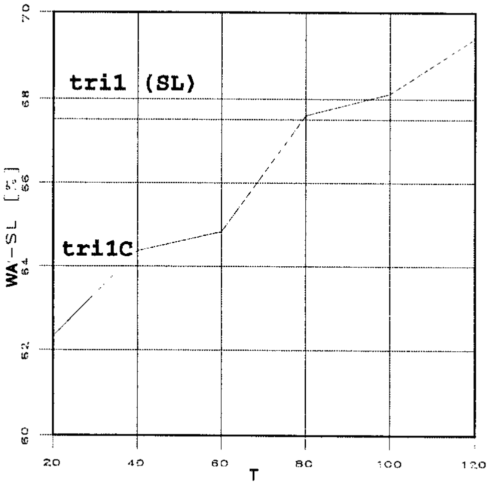


Figure 1. Word accuracy in case of monolingual experiments (Slovene language) using the multilingual set of triphones produced at various threshold values.

Next the value of weight *C* was increased to 0.5 (actual values of weights was *L,C, R was 2, 1, 2,* respectively, since only integer values were allowed). Increasing the value of weight *C* was found to improve the performance of the recogniser with multilingual set of triphones (Table 3.b). The WA and LI rates were similar as for the *C=0*, however the *GCR* was much higher (Figure 2). In this case the the *GCR* of 54.57% was achieved at approximately 1.6% decrease of *WA* rate and less than 5% decrease of *LI* rate. Such reduction of the total number of triphones with minor degradation of the *WA* can be considered as an improvement of the baseline system performance. As for the case of *C=0,* the use of multilingual set of triphones for the monolingual experiments improved the word accuracy in case of Slovene language for the threshold values of 400 or more.

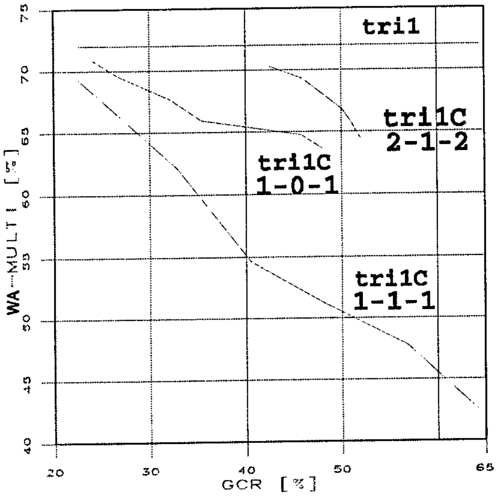


Figure 2. Word accuracy in case of multilingual experiments for various values of weights *L,C* and *R* as a function of *GCR*.

Further increase of weight *C* (*C=1*) did not improve the performance of the recogniser with multilingual set of triphones (Table 3.c). Setting the *C* to 1 can produce the multilingual set of triphones with the highest *GCR,* but on the other hand, it significantly reduces the *WA* and *LI* (Figure 2).

a) *L=1,C=0,R=1*

models	T	N	WA				LI	GCR
			SL	ES	DE	MULTI		
tri1C	20	6498	62.34%	65.92%	69.51%	63.68%	75.73%	47.99%
tri1C	40	6799	64.38%	67.34%	70.73%	64.74%	77.57%	45.79%
tri1C	60	8226	64.83%	68.23%	72.23%	65.94%	78.23%	35.38%
tri1C	80	8662	67.60%	69.81%	73.90%	67.63%	80.37%	32.19%
tri1C	100	9424	68.12%	70.10%	74.95%	69.57%	84.52%	26.63%
tri1C	120	9758	69.41%	72.67%	75.85%	70.84%	86.07%	24.19%
tri1	-	13074	68.04%	78.58%	76.77%	71.99%	91.61%	0%

b) *L=2,C=1,R=2*

models	T	N	WA				LI	GCR
			SL	ES	DE	MULTI		
tri1C	100	5942	63.36%	63.21%	72.11%	64.43%	78.95%	52.04%
tri1C	160	6026	65.29%	64.02%	73.57%	65.14%	79.22%	51.43%
tri1C	180	6068	66.40%	64.89%	74.81%	65.63%	79.75%	51.12%
tri1C	260	6208	66.91%	66.12%	75.98%	66.82%	81.21%	50.10%
tri1C	340	6784	67.73%	69.47%	76.51%	69.27%	84.78%	45.89%
tri1C	400	7239	69.23%	73.20%	76.68%	70.32%	86.67%	42.57%
tri1	-	13074	68.04%	78.58%	76.77%	71.99%	91.61%	0%

c) *L=1,C=1,R=1*

models	T	N	WA				LI	GCR
			SL	ES	DE	MULTI		
tri1C	120	4238	29.12%	38.95%	40.72%	42.35%	58.89%	64.47%
tri1C	140	5284	28.63%	45.81%	47.34%	47.83%	63.55%	56.84%
tri1C	180	6475	37.78%	52.73%	53.59%	51.21%	69.26%	48.15%
tri1C	200	7526	46.62%	57.37%	59.45%	54.67%	76.31%	40.48%
tri1C	240	8577	56.16%	62.48%	64.83%	62.13%	82.74%	32.81%
tri1C	280	9971	65.41%	69.41%	71.73%	69.25%	86.07%	22.64%
tri1	-	13074	68.04%	78.58%	76.77%	71.99%	91.61%	0%

Table 3. Performance of the recogniser using various multilingual sets of triphones produced at different values of weights *L,C,R..*

7. CONCLUSION AND FUTURE WORK

Experiments have shown that the use of clustering algorithm can produce the multilingual set of triphones that achieves almost the same word accuracy as the language specific triphone sets operating in parallel. Slight decrease of the word accuracy is acceptable considering the fact that the number of triphones in a multilingual set of triphones is significantly smaller than total number of triphones in the language specific triphone sets. In best case the use of clustering algorithm resulted in a reduction of the number of triphones by more than 40% with degradation of word accuracy by 1.67%: Such result shows that the multilingual set of triphones produced by the clustering algorithm can improve the performance of a multilingual recogniser based on language specific triphone sets operating in parallel.

The monolingual experiments with multilingual set of triphones have shown that in some cases the use of multilingual set of triphones can also improve the performance of the monolingual recognisers, that is, the performance of the recognisers based on monolingual triphone sets. Such improvement has been observed for the Slovenian language where the performance of the recogniser using the Slovenian triphone set was significantly lower than the performance of the recogniser (based on the Spanish and German triphone sets) for the Spanish and German languages. The multilingual set of triphones tends to equalise the performance of all monolingual triphone sets that were used for definition of the multilingual triphone set.

Results of the monolingual experiments using the multilingual triphone set indicates that the multilingual triphone set might also perform well for the new languages, that is the languages that were not included during the definition of the multilingual triphone set. However, no experiments have been done so far to prove this.

In future, the number of SpeechDat databases will be increased in order to expand the scale of experiments and to provide more reliable assessment of the clustering algorithm efficiency.

The clustering algorithm bases on a definition of distance measure for triphones defined as a weighted sum of explicit estimates of the context similarity on a monophone level. In this case the monophone distance estimation method was based on the algorithm of Houtgast. In future, other methods of monophone distance estimation will be also considered.

8. REFERENCES

[1] O. Andersen, P. Dalsgaard and W. Barry, *Data-Driven Identification of Poly- and Mono-phonemes for four European Languages.* 1993, Proc. EUROSPEECH '93, Berlin, pp. 759 - 762

[2] J. Koehler, *Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds.* 1996, Proc. ICSLP '96, Philadelphia, pp. 1780 - 1783

[3] K. M. Berkling, *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters.* Oregon Graduate Institute of Science & Technology, Dissertation, 1996

[4] P. Bonaventura, F. Gallocchio, G. Micca, *Multilingual Speech Recognition for Flexible Vocabularies.* 1997, Proc. Eurospeech '97, Rhodos

[5] F.Weng and H. Bratt and L. Neumeyer and A. *Stolcke, A study of Multilingual Speech Recognition.* 1997, Proc. Eurospeech '97, Rhodos

[6] JanezKaiser, Zdravko Kačič, *Development of the Slovenian SpeechDat.* Speech Database Development for Central and Eastern European Languages, Granada,1998

[7] H. Hoege, H. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach, *European speech Databases for Telephone Applications.* 1997, Proc. ICASSP '97, Muenchen, pp. 1771 - 1774

[8] A. Haunstein, E. Marschall, *Methods for Improved Speech Recognition over the Telephone Lines.* Proceed. IEEE IC ASSP, 1999

SPEECH RECOGNITION OF NON-NATIVE SPEECH USING NATIVE AND NON-NATIVE ACOUSTIC MODELS

David A. van Leeuwen and Rosemary Orr

vanLeeuwen@tm.tno.nl
TNO Human Factors Research Institute.
Postbus 23,
3769 ZG Soesterberg,
The Netherlands.

R.Orr@kno.azn.nl
University Hospital Nijmegen St Radboud
Philips van Leydenlaan 15
5600 HB Nijmegen
The Netherlands

ABSTRACT

A speech recognition system is subjected to the speech of non-native speakers, using both native and non-native acoustic phone models. The problems involved with the mapping of phoneset from the non-native to native language are investigated, and a detailed analysis of phone confusions is made. For Dutch speakers, British English acoustic models give the best word recognition results.

INTRODUCTION

The study of speech as uttered by non-native speakers of the language has been a subject of research in phonetics for along time [1]. With maturing speech technology, the subject of non-native speech is becoming a topic of interest. Non-native speakers will form a new challenge for any technology for which acoustic training is an important factor, *e.g.*, codebook based coding systems, or speech and language recognition systems. One of the problems of training for non-native speakers is that the target group is very inhomogeneous—there are in principle as many potential non-native classes as there are languages in the world. This is a larger number than the number of dialects within a language, which has been the classic acoustic modelling challenge.

The standard approach for a technology such as speech recognition is to gather a database of the target group of users, and (re)train the system using this new database. For non-native speech, this means that if there are N languages for which speech recognition is available, a full matrix of native and non-native recognition systems would require N^2 speech databases, most of which will be non-native databases. Currently, the number of available non-native databases is very limited.

An alternative approach to non-native speech is to assume that non-native speakers will dominantly use their native phones, presumably by mapping the phones of the language they are speaking (L2) to their native language (L1). If this is the case—and the fact that foreign speakers can very well be characterized (and caricatured) supports this assumption—a speech recognition system can use the L1 phone models for the non-native speakers, combined with L2 dictionary and language models. In this way, only N acoustic training databases must be available for a full set of native and non-native recognition systems in N languages. Of course, there are non-native issues in pronunciation rules (dictionary) and language modelling as well, but we will not address these in this study.

This paper reports on an experiment for Dutch speakers speaking English, where a speech recognition system is trained with either native Dutch, British English or American English speakers. The main objective is to investigate whether L1 speakers should be recognized using L1 or L2 acoustic models when they are speaking in a non-native language L2. The implementation is limited—only one non-native language combination is investigated, using only one speech recognition system—and therefore the methodology of the experiment might have more implications than the bare results.

THE MIST SPEECH DATABASE

In late 1996 TNO recorded a speech database for Dutch continuous speech recognition named NRC0, similar to the Wall Street Journal corpus, (WSJ0) [2]. The main purpose of the database was to bootstrap the development of large vocabulary continuous speech recognition for the Dutch language. This database consisted of 52 speakers, each uttering 65 unique sentences. The sentence texts were taken from a Dutch newspaper (*NRC/Handelsblad*), read from a CRT screen in a quiet and low reverberant room, using a Sennheiser HMD 414-6 microphone, and high quality digital recording equipment. The number of speakers is smaller than for WSJ0 (and similar databases such as WSJCAM0 [3] and BREF80 [4]), and therefore TNO decided to extend the database in 1998 with another 80 speakers (NRC1). For these sessions, special sentences were recorded additional to the 65 utterances for continuous speech recognition systems. These included 'foreign language sentences,' which were sentences in English, French and German. The prompt texts for the foreign language sentences were taken from newspaper texts, English from *Wall Street Journal*, German from *Frankfurter Rundschau* and French from *Le Monde.* These were the same sources from which the development and test sentences in the SQALE project [5] were chosen. The recordings of the foreign language sentences can be considered non-native speech material for English, French, and German.

The majority of the speakers for NRC1 were recruited from the institute. Of the institute's employees, 60 % has an academic background, and 20 % a higher technical education. This is not a representative sample of the Dutch population. There is the advantage, however, that a relatively high fraction of subjects can be expected to be able to speak one or more foreign languages. It was left to the subject's own discretion to decide whether or not to record the foreign sentences. Thus, of the 74 subjects that recorded foreign speech, 71 recorded English, 66 German and 60 French. The prompt texts consisted of five sentences that were the same for all speakers, and could function as adaptation sentences. A further five sentences were chosen, which were unique for every speaker.†

For the purpose of the MIST workshop, TNO decided to share the non-native speech data with other research institutions. A liberal license agreement allows people to use the speech material for research purposes, free of charge. As a reference, 10 Dutch sentences per speaker were added to the non-native speech database, again consisting of 5 sentences that were the same across all speakers, and 5 unique sentences. Thus a total of over 5 hours of speech is available for the scientific community. Only for the Dutch sentences, a detailed orthographic transcription could be made, for the other three languages just the prompt texts were distributed. It is hoped that native speakers at other institutes will provide the community with corrected transcriptions.‡ A number of articles in these proceedings [6, 7] already report on results using this database. For the experiments in this paper, only the English utterances were used.

THE ABBOT SPEECH RECOGNIZER

For the speech recognition system used in this experiment, we used the Abbot large vocabulary continuous speech recognition system [8]. Abbot is a hybrid neural net/Markov model recognition system. The most important difference from traditional hidden Markov model systems is that the neural net directly estimates *a posteriori* phone probabilities for each speech frame. The forward pass in the recurrent neural net can be calculated quickly, and phone probabilities are quite well determined. This makes the decoding search relatively easy, and therefore the system is known for its fast recognition speed. By choosing the appropriate decoder, both a phone recognition system and a word recognition system can be built.

The components needed for the various word recognizers are

- L2 (English) and L1 (Dutch) acoustic models
- L2 dictionary
- L2 to L1 phoneset mapping
- L2 language model.

When Dutch acoustic models are used, a dictionary of English words in terms of Dutch phones is needed. One way to achieve this is to use an English dictionary, and to translate all English phones into corresponding Dutch phones. For this the reason the L2 to L1 phoneset mapping is necessary.

For a phone recognizer, the phone mappings appear to be unnecessary. However, for evaluation of the Phone Error Rate (PER) a phone level reference transcription is needed. Because the test database is annotated at the word level, a dictionary is needed to convert the L2 reference words into L1 phones. As English dictionaries in terms of Dutch phone sets are not available, the phone mapping is necessary in this case as well.

EXPERIMENTAL SETUP

The test database used is the English part of the MIST speech database. The speakers were separated into two groups, training and testing speakers. The training speakers were not used in this experiment, and only the five unique sentences per speaker were used. This resulted in 180 utterances by 36 speakers. Of the 3147 words 129 (4 %) words were Out Of Vocabulary (see below).

Table I. Acoustical training conditions for three languages.

Language	American	British	Dutch
Database	WSJ0	WSJCAM0	NRC0
# speakers	84	90	48
speech length (hr)	13	13	7
phones	53	44	39
phoneset	ICSI/LIMSI	BEEP	CELEX

Three different acoustical models were used, American English, British English, and Dutch. The training conditions are comparable, except for Dutch, which has about half the training time (7 hr). The Abbot speech recognition system is known to have a relatively quickly saturating performance with increasing training data, due to the limited number of parameters to be estimated. In table I the acoustical conditions are tabulated. The phoneset for Dutch is a subset of the phoneset defined in the CELEX dictionary [9]. For American English, the ICSI/LIMSI phoneset is used [10, 11]. The training for American and British English was performed by Cambridge

† For each language, there are 2–5 sentences that occur twice among the speakers, due to an unfortunate misconfiguration during the sentence selection.

‡ The latest transcriptions can always be found at URL `ftp://ftp.tm.tno.nl/pub/speech/mist`.

Table II. The phone map used in order to translate the American and British English dictionaries using the Dutch phoneset. The second and fourth conlumn show the full English phonesets, the middle column shows the Dutch phones to which the phones are mapped. The phones f, h, ʤ, l, m, n, ŋ, s, ʃ, v, j, z, ʒ occur in all three phone sets, and are not shown.

American→		Dutch	←British	
bottle	ɑ	ɑ	ɑ	heart
hamm	æ	ɛ	æ	zap
sum	ʌ	ɑ	ʌ	rough
might	aɪ	aː j	aɪ	ice
more	ɔ	ɔ	ɔ	lord
		ɔ	ɒ	pot
ago	oʊ	oː	əʊ	rogue
annoyed	ɔɪ	ɔ j	ɔɪ	boil
house	aʊ	aʊ	aʊ	house
again	ə	ə	ə	again
alive	l̩	ə l		
atom	m̩	ə m		
heaven	n̩	ə n		
after	ɹ̩	ə ʀ		
hurd	ɝ	ə ʀ	ɜː	burn
bet	ɛ	ɛ	ɛ	bet
		ɛ ə	ɛə	hair
pain	eɪ	eː	eɪ	pain
adding	ɨ	ɪ		
fit	ɪ	ɪ	ɪ	fit
		ɪ ə	ɪə	here
beat	i	iː	i	beat
hook	ʊ	uː	ʊ	bush
cool	u	uː	u	cool
		uː ə	ʊə	poor
lobe	b̚	b	b	board
bow	b			
beach	ʧ	t ʃ	ʧ	beach
shed	d̚	d	d	does
does	d			
this	ð	d	ð	that
butter	ɾ	d		
jig	g̚	g	g	go
go	g			
aha	ɦ	h		
arc	k̚	k	k	cow
cow	k			
chip	p̚	p	p	pot
pot	p			
raise	ɹ	ʀ	ɹ	raise
fit	t̚	t	t	tip
tip	t			
thing	θ	t	θ	thing
walk	w	ʋ	w	walk

University. In the American English training procedure, a different dictionary was used [12].

The size of the vocabulary was conservatively chosen to be 20k words. The limited size was used because it was not an objective to optimize a system for performance, but rather to compare performances. The vocabulary and dictionaries were effectively determined by the freely available demonstration version of Abbot [13]. The American English pronunciation dictionary is based on the CMU dictionary [14], whereby the phoneset was converted using an automatic phone mapping to the ICSI phone set. The British English dictionary is a subset of the BEEP dictionary [15]. In order to obtain dictionaries for the Dutch phone set, both dictionaries were translated using a phone map shown in table II.

The language model used is a 20k word trigram language model, which was developed using American English texts pre-dating spring 1998. The decoder used for Abbot is 'chronos,' a time-synchronous stack decoder [16]. The language model was used for all word recognition runs, except for the Dutch baseline run.

Phone mapping

The phone mapping shown in table II needs some explanation. It was based on our phonetic intuition of the similarity between Dutch and English phones. The table shows only one mapping per phone, but later we will show that experiments have been carried out with multiple mappings. Some phone mappings have been made consistent with the way the Dutch vocabulary, that was used in the acoustic model training, expresses words in terms of the Dutch phones. For instance, the mapping [aɪ] → [aː j] is chosen over [aː ɪ], because the CELEX dictionary has entries for words like *haai* → [h aː j] (shark). In the training process of the Dutch acoustic models, therefore, the [j] models the [ɪ] in the context of [aɪ].

The American English phoneset has separate entries for 'closures,' plosives without an audible release, [b̚, d̚, g̚, k̚, p̚, t̚], in combination with the standard IPA plosives. In the dictionary used for American English, most occurrences of a plosive are preceded by a closure, *e.g.*, *bee* → [b̚bi]. However, the non-audible release can stand on its own, *e.g.*, as in *add* → [ad]. For this reason, the closures are mapped to Dutch plosives, and the plosives are mapped to nothing.

BASELINE RESULTS

In order to have a reference for the experiments with non-native speech, a number of baseline tests were performed. For this, development test material used in the SQALE project was used. This consisted of 20 native speakers for American and British English (10 male, 10 female). For a baseline for the Dutch models, 20 speakers of the NRC1 Dutch database were used. Each of the speakers contributes 10 utterances to the test. In table III the phone and word errors of the recognizer are given. The baseline results are only indicative of the recognition system; they are not 'optimal' values. For instance, the language model has not been optimized for the speech domain. In determining the PER for English, an automatic expansion of the reference word transcriptions has been made, using the appropriate dictionary. Because the English dictionaries have multiple pronunciations per word, many arbitrary decisions have been made in generating the phone reference transcription. This leads to an estimation for the PER which is too high.

The Dutch dictionary has only single pronunciation entries. This may be the reason that the PER figure is lower than for English. The word error rate for Dutch is much higher than for English. The Dutch language model was based on a 78 million words text of newspaper text, defining the vocabulary as the most frequent 20 000 words. The language model was built specifically for the baseline test, and has not been optimized.

Table III. Word and phone error rates (WER) in % for baseline conditions. The top line gives the WER in the standard 'forward' condition, that is used throughout this work. 'Forward' means a forward pass only, 'fw/bw' means forward and backward pass (see text).

Language	American	English	Dutch
WER (forward)	**27.6**	**26.2**	**37.7**
WER (fw/bw)	22.4	22.5	34.4
PER (forward)	39.8	37.4	35.6
PER (fw/bw)	37.3	34.7	33.4

In table III results for a forward/backward pass are given as an indication as to how much lower the error rates are if the posterior log probabilities are averaged with 'backward' runs. Because Abbot utilizes a recurrent neural network, past acoustic context is automatically modelled. In order to model future acoustic context, a 'backwards' network can be trained by feeding the network acoustic features that are reversed in time. In the recognition pass, the backwards classified phone probabilities can be merged with the forward stream, which generally leads to lower error rates.

Results for non-native speech

In table IV the results for the MIST database are given. Results obtained with Dutch models are made with either US or UK dictionary, translated using the phone map of table II.

Table IV. The word and phone error rates (in %) for the MIST database of Dutch speakers speaking English. A 20k English vocabulary and an accompanying trigram language model was used. (US, UK, NL) means American, British, Dutch. The standard deviation of the numbers is approximately 0.8 %.

Acoustic models	US	UK	NL	NL
Pronunciation dictionary	US	UK	US	UK
WER	68.8	60.9	68.9	73.4
PER	55.7	49.1	54.5	56.2

It appears that British acoustic phone models give the lowest error rates for the Dutch MIST speakers, both in phone and word error rate. It is interesting to note, that the difference between PER and WER is smaller—and has actually reversed sign—with respect to the baseline. One is tempted to assign this to a language model incompatibility, but this is unlikely because of the very similar source of both tests, namely the SQALE sentences.

Influence of the phone mapping

Of the results of table IV, the last two columns are most interesting, because they involve a non standard combination of L1 acoustic models and L2 language models. The phone mapping shown in Table II is the first mapping we tried, based on phonetic intuition. The Dutch phoneset contains some phones that are not covered by the English mappings, namely [øː, ei, ʉ, œy, x], and [yː]. Other English phones, that experienced speakers are capable of using, have no real Dutch equivalent. Examples of these are the infamous 'th' consonants [θ] and [ð]. We experimented with a couple of changes to the phone mapping, in order to investigate if any of them would lower the word error rate.

First, we adapted the dictionary conversion tool to accept alternatives for phone conversions. This involved a recursive expansion of alternative pronunciation strings. For instance, if both the alternatives [ð] → [d|z] and [ʌ] → [ɑ|ʉ] are allowed, the word 'mother' ([mʌðɹ̩]) gets four alternative pronunciations, [mɑdəʀ, mʉdəʀ, mɑzəʀ] and [mʉzəʀ]. The inclusion of the above examples and [θ] → [t|s] lead to an *increase* in word error rate of 7 %-point for the American English dictionary. Apparently, allowing more pronunciation variants per word causes more options for erroneous words than that it helps to find options for the correct word.

We have run several tests in order to investigate what the individual contribution of the alternatives to this increase is. The alternatives that we defined for American and British pronunciation are shown in the first columns of table V, together with the difference in WER the individual alternative makes. Again, almost all alternatives lead to an *increase* in word error rate.

Table V. Changes from the default phone mapping (see table II). In the last column, the increase in the word error rate (in %-point) with respect to the baseline is given.

English	Dutch	US	UK
ʌ	ɑ|ʉ	+2.6	+1.4
ʌ	ʉ	+4.3	+3.0
θ	t|s	+0.5	+2.1
θ	s	+0.8	+2.1
ð	d|z	+2.9	+1.8
ɾ	d|t	−0.8	
ɾ	t	0.0	
aɪ	aː iː	+2.2	+3.4
aɪ	aː ɪ	+2.0	+2.3
ɨ	ə	+0.8	
ɛə	ɛ ʀ		−0.2
ʊə	uː ʀ		−0.5

The increase of PER for the mapping [ʌ] → [ʉ] surprised us, because in the stereotypical Dutch En-

Table VI. Individual phone confusions. Only phones that are confused more often with others (left number) than that they are recognized correctly (right number) are shown. The leftmost columns show the phone confusion considered. The second three columns show the phone confusion numbers for the baseline tests. The third and fourth three columns show the confusion numbers for the non-native database. Boldface indicates more errors than correct. In the case of the Dutch phoneset (last group of rows), a dictionary phone mapping for the reference transcription was used.

Phones			Reference			Non native database MIST					
Set	ref.	rec.	test	err	corr	mapping	err	corr	mapping	err	corr
US	m̩	m	SQALE	**27**	2	US	**29**	1			
	n̩	n		**163**	42		**332**	56			
	ɦ	h		**2**	1		**5**	1			
	ʒ	ʃ		0	6		3	3			
	ɨ	ɪ		**146**	74		**207**	53			
UK	a	ɛ	SQALE	9	126	UK	**112**	55			
	ʒ	ʃ		0	6		**2**	1			
NL	ɑ	aː	NRC1	141	2679	UK	**184**	166	US	74	183
	ɔ	oː		254	1580		223	340		**181**	167
	ʊ	r		3	225		**36**	7		0	0
	ʒ	ʃ		**15**	0		**2**	0		**3**	0
	ʤ	j		**6**	0		**18**	0		**19**	0
	f	v		218	517		**175**	161		**178**	161
	g	x		4	4		**18**	4		**9**	5
	g	k		**11**	4		**80**	4		**81**	5
	v	f		303	1968		**216**	164		**214**	162
	z	s		214	1221		**268**	149		151	266

glish accent [ʌ] is pronounced as [ʉ]. The reason might be, that the acoustic modelling for [ʉ] in Dutch is relatively poor. The confusibility of [ʉ] with [ə] is high because the schwa lies acoustically very close to the unstressed [ʉ].

One more elaborate expansion is that of the plosives in the American English phone set. The mapping of [b, d, g, k, p, t] to nothing leads to a few errors in the converted dictionary. Words like *update* have the US expansion [ʌp˺deɪt˺], where there is a closure of /p/ followed by the release /d/. In our original mapping, the latter phone was deleted. Correcting for these occurrences (translating *update* → [ɑpdeːt]) lead to a *decrease* of the word error rate for the American dictionary of 0.3 %-point. A combination of this with the alternative [ɾ] → [d|t] lead to a total decrease of 1.2 %-point.

Individual phone scores

By investigating the phone recognition result, it is possible to make an inventory of the individual phone scores. A phone class based alignment algorithm [17] can provide a fairly good measurement of the phone confusion matrix, even for continuous speech recognition. A way to summarize the problems in phone recognition is to tabulate the phones that are recognized more often as a different phone than as themselves. In table VI these phones are indicated for a number of baseline and non-native tests.

In some cases we can conclude that the basic models are not well trained. This is the case for, *e.g.*, the American [m̩, n̩] and [ɨ], and the Dutch [ʒ, ʤ] and [g]. But for other phones, there is a clear effect of the non-native speech. From table VI it is clear that the British [a] is pronounced closer to the [ɛ] by the Dutch speakers. When the Dutch phoneset is used for the non-native speaker, there are many examples of phones that have a high confusibility with others. This may be an artifact of the automatic dictionary mapping. Interestingly enough, both [f] and [v] have a tendency to be interchanged in recognition with respect to the dictionary expansion. Our understanding of this is that in Dutch local accents, the /f/ and /v/ have acoustic realizations that are similar, because the difference in voicing tends to blur.

CONCLUSIONS

We have shown a methodology that allows non-native (L2) speech recognition using native (L1) speech models, L2 dictionary and grammar, and an L2 → L1 phone mapping. In the case of Dutch non-native speakers of English, the plain word recognizer using British English models gives lower word error rates than the approach given above, but it is not known whether this will generalize to other combinations of non-native speech. Still, the word error rate of the non-native speakers is a factor 2 higher than for native speakers. The phone mapping, necessary in order to define a L2 dictionary in terms of L1 phones, forms a weak link in the approach. A more elaborated rule based translation of the vocabulary should lead to better results for the approach taken here.

REFERENCES

[1] James Emil Flege, Ocke-Schwen Bohn, and Sunyoung Jang. Effects of experience on non-native speakers' production and perception of english vowels. *Journal of Phonetics*, 25:437–470, 1997.

[2] D. B. Paul and J. M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Proc. Fifth DARPA Speech and Natural Language Workshop*, pages 357–362. Morgan Kaufmann Publishers, Inc., 1992.

[3] Jeroen Fransen, David Pye, Tony Robinson, Phil Woodland, and Steve Young. WSJCAM0 corpus and recording description. CD-ROM documentation, 1994. CUED Cambridge (UK).

[4] L. F. Lamel, J. L. Gauvain, and M. Eskenazi. BREF, a large vocabulary spoken corpus for French. In *Proc. Eurospeech*, 1991.

[5] S. J. Young, M. Adda-Dekker, X. Aubert, C. Dugast, J.-L. Gauvain, D. J. Kershaw, L. Lamel, D. A. van Leeuwen, D. Pye, A. J. Robinson, H. J. M. Steeneken, and P. C. Woodland. Mutilingual large vocabulary speech recognition: the European SQALE project. *Computer Speech and Language*, 11:73–89, 1997.

[6] R. Wanneroy, E. Bilinski, C. Barras, M. Adda-Decker, and E. Geoffrois. Acousic-phonetic modeling of speech for language identification. In *These Proceedings*, pages 9–13, 1999.

[7] Geoffrey Durou. Multilingual text-independent speaker identification. In *These Proceedings*, pages 115–118, 1999.

[8] Tony Robinson, Mike Hochberg, and Steve Renals. *The use of recurrent networks in continuous speech recognition*, chapter 7, pages 233–258. Kluwer Academic Publishers, 1996.

[9] R. H. Baayen, R. Piepenbrock, and L. Gulikers. The CELEX lexical database. CDROM, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.

[10] J. L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker. Speaker-independent continuous speech dictation. *Speech Comm.*, 15:21–37, 1994.

[11] D. J. Kershaw. *Phonetic Context-Dependency in a Hybrid ANN/HMM Speech Recognition System.* PhD thesis, University of Cambridge, January 1997. URL: `http://www-svr.eng.cam.ac.uk/~djk/Publications/thesis.html`.

[12] Tony Robinson. Private Communication.

[13] See URL. `http://svr-www.eng.cam.ac.uk/~ajr/abbot.html`.

[14] See URL. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`.

[15] See URL. `ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries`.

[16] Tony Robinson and James Christie. Time-first search for large vocabulary speech recognition. In *ICASSP*, 1998.

[17] David A. van Leeuwen and Michael de Louwere. Objective and subjective evaluation of the acoustic models of a continuous speech recognition system. In *Proc. Eurospeech*, pages 1915–1918, 1999.

Report of the plenary discussion on "Non-native speech and Accents"

Chairperson: Denis Johnston (BT, United Kingdom)
Reporter: Els den Os (KPN-Research, the Netherlands)

Questions from *Hunt* and *Boves*: Both commented on the assumption made by Van Compernolle that accents can be considered as "information preserving transforms," in not more homophones in one accent than in another. Van Compernolle also argued that speech recognisers should impose constraints to exploit this consistency. There are a number of issues that call these assumptions into question.

1 Accents are not only related to the sound structure in a language but also to the vocabulary

2 French provides a counter example. For many (young) speakers pairs of phonemes are becoming collapsed into single phonemes, often with context dependent allophones corresponding to the earlier phoneme pair. Pairs that are collapsing: /e/–/ɛ/ et–est; /a/–/ɑ/ pâte–patte. In some Dutch accents the distinction between voiced and voiceless fricatives has disappeared.

3 Since people move around more in their lives and as they are increasingly influenced by telecommunications, accents will increasingly become inconsistent. Imposing consistency constraints on the recogniser could then be harmful rather than helpful. On the other hand, mobility and telecommunication may cause accents to become more similar. This should help ASR.

Reaction *Van Compernolle*: There is indeed no direct evidence for considering accents in this way, however, it is a good working hypothesis.

Comment from *Hunt*: In a reaction on the presentation by *Geoffrois*, *Hunt* mentioned an explanation for the fact that language identification was better for French natives than the identification of English and German for the native speakers. This explanation relates to the fact that French has relatively few consonant clusters, while English and German share the property of having many consonant clusters.

Reaction *Geoffrois*: This may well be the case. When looking at the confusion matrix it was noticed that there were more confusions between English and German than between French and either of the two other languages.

Van Leeuwen pointed out that there might be an explanation for the fact that no language dependency results were found for the non-natives, while this was the case for natives. This explanation is related to the fact that the same group of Dutch speakers produced the English, German and French items, while the native items were produced by different groups of speakers. *Hunt* indicated that another explanation might be that Dutch speakers speak better German and English than French.

Reaction *Geoffrois*: This might well be the case. The results in the paper of Geoffrois/Durou also point into this direction.

Reaction *Boves*: It might be worthwhile to look at the output of the decoder to see the succession of phones, since the bigram training is done on the decoder output.

Questions to *Eklund*: Some questions were related to the collection and transcription of the speech data on which his study is based. *Van Compernolle* indicated that the type of instructions may have an influence pronunciations. *Eklund* pointed out that the speakers did not receive any special instructions on how to pronounce the names in the sentences. Furthermore, *Eklund* explained that he chose proper names only, because only in this way you can obtain many different non-native sounds. There has been no check of the level of agreement between the three labellers.

Reaction *Van Compernolle*: Xenophones† should be used for training. This gives at least some information even if there are not enough data. There is always the risk that there are too few different words with non-native sounds so that the xenophone models become context dependent

Reaction *Hunt*: referring to earlier work, adding two Scottish phonemes to English improved the performance of at least one recognizer substantially.

Reaction *Van Leeuwen*: it is really very important to properly choose your phone set.

General comments, not related to a specific paper

Boves mentioned that we should really ask ourselves: what is the nature of the models of todays ASR systems. An important issue is the distinction between the discrete and symbolic representation (*e.g.*, IPA) on the one hand and the continuity of speech on the other hand.

Related to this, he suggested that foreign accent might be primarily in the dynamics, especially for languages that are phonetically close. The question is how to model this dynamics and how and where does this interfere with ASR.

Hunt indicated that the work of Li Deng and John Bridle on Hidden Dynamic Modelling is interesting in this respect. It was also mentioned that we should try to find more robust phonetic units.

Two general issues were mentioned related to non-native speech:

1. The need for the collection of more non-native databases
2. How to deal with unseen data

Related to point 1: *Schulz* indicated that only collecting extra databases would not solve the problem, because of the diversity of non-nativeness. We should try to find rules to deal with non-nativeness. However, it is questionable whether these rules are easy to derive.

Micca said that we should get better ideas on when it is better to have pronunciation variants in the lexicon and when it is better to have non-native acoustic models.

Adda-Decker mentioned that it might be a very good idea not only to look at the recognition errors, but also to look at what is correctly recognised.

Related to point 2: *Van Compernolle* indicated that we must find ways to deal with unseen data. The range of variation due to non-nativeness is too large to hope that we can build models from just more data. The question is how to get the best models. Shifting means of distributions is a possible solution, provided that we know how to do this correctly. This corroborates the opinion expressed by *Schulz.*

† Xenonphones are phones that approximate phones from other languages (see paper from Lindstrom and Eklund)

AN OVERVIEW OF THE EURESCOM MIVA PROJECT

Denis Johnston

BT Adastral Park, Ipswich IP5 3RE ,England,UK

e-mail: denis.johnston@bt.co.uk

Phone: +44 1473 64 2128

ABSTRACT

The goal of the MIVA project was to answer a number of fundamental questions concerned with the exploitation of speech technology enabled systems. The experimental service chosen was designed to help foreign people travelling in the country to find emergency and embassy numbers, country and area codes, useful numbers (directory service, country direct, etc.) and how to use Telecom and credit cards for placing calls.

Services were implemented in each of the countries taking part and two stages of experimentation were undertaken. The first of these was a mono-lingual experiment carried out in each country to optimise performance for each country /language combination. The second was a fully multi-lingual service in which each of these optimised services was re-implemented in all languages. All systems were evaluated over combinations of local and international environments. Correlations derived from a subset of the subjective and objective results were used to provide a predictive model of users opinions and the remaining subset of data used to test these predictions.

1. INTRODUCTION

Many services based upon advanced speech technology such as ASR have been implemented in recent years but in the main these services have been limited to one language. In this project we undertook basic research into how such systems might be implemented, applied and evaluated in a multi-lingual environment. With so many interacting factors likely to impact upon users' perceptions selecting the best combination is far from trivial. Previous experience, recogniser accuracy, recogniser speed, recogniser threshold settings, vocabulary choice, characteristics of spoken prompts, line types, phone types, dialogue characteristics are all important and interact strongly. Simultaneously optimising them can be exceedingly difficult. Dealing with is complexity was expected to be a major challenge. This multi-dimensional problem has generally resulted in two broad approaches to evaluation. One has been to arbitrarily choose one of the more obvious and directly measurable factors such as recogniser threshold settings, attempt to hold all others constant and then measure some other quantifiable effect such as total transaction time. The second has been to set up the service, invite people to use it and then use questionnaires or interviews to collect opinions about the quality or usability of the service.

However neither of these is really satisfactory. The problem with the first is that there are few factors that are directly measurable and meaningful. More often it is the unmeasurable and intangible features of the service such voice quality and dialogue styles that tend to dominate user perceptions. On the other hand in the field trial approach, reliable data collection (especially of users perceptions) is difficult and expensive. And subsequent analysis and interpretation of the results in what is an almost totally uncontrolled environment is often impossible.

In the MIVA project we substantially circumvented these problems by adopting a methodology based upon multi-factor experimental designs. This approach has been widely applied in other disciplines and is one of the standard methodologies applied in the life sciences[1]. We show how this approach overcome all of the disadvantages of the methods described above and allowed decisions concerning the best 'mix' of characteristics to be determined.

2. ASKING THE RIGHT QUESTIONS

2.1 High level questions

The starting point for such a process is a list of high level questions. This was done in consultation with the technical and marketing departments of our organisations. This ensured a high level customer led drive for the project and simplified the agreement on what issues were of importance to all concerned. Once the questions were established they were prioritised.

The key questions identified were:

- Do users prefer speech recognition to DTMF input for Interactive Voice based services?
- Is the performance of speech recognisers significantly impaired when services are accessed over GSM networks?
- Are there certain preferred dialogue structures ?
- What speech recogniser parameters (e.g. rejection settings, cut- through strategies) are preferred?
- What is the best way to prompt users at the start of a dialogue so that their language can be identified?
- Do recognisers perform differently when accessed over international links?

In examining these questions it became apparent that they could best be answered using two separate series of experiments. The first series would be undertaken independently in each country and would address those questions (numbers 1-3) that did not have a direct multi-lingual dimension. This first phase would also be used to optimise platforms and dialogues and identify the appropriate prompts and words to be recognised. The second phase would embrace those tests that demanded multi-linguality.

2.2.Supplementary questions

During the course of project it became apparent that we could and should, address other questions such as:

How should the problem of 'unanswerable' queries be resolved? For example if a person accesses an information system expecting to obtain information about telephone fault repair how do you deal with the problem that this information is not in the database?
What, if any, are the important differences, due to networks, country size, languages, cultures etc. which must be taken into account when providing multilingual services?

2.3 From questions to hypotheses

Once the questions had been established we were able to move to the experimental design stage. The first step in that direction was to convert the questions into the 'null hypothesis' format necessary to allow statistical tests to be performed at the subsequent analysis stage.

Formulations of null hypotheses for the above questions are:

- Users show no preference or behavioural patterns between speech recognition or DTMF
- The performance of speech recognisers is not impaired when services are accessed over GSM networks?
- All dialogue structures are equally efficient and effective.
- Recogniser parameters such as rejection settings, cut- through strategies make no difference to performance or user behaviours.

The value of recasting the questions into a null hypothesis format comes from the fact that the onus shifts from having to prove something true to proving it untrue.

2.4 Selecting an appropriate service

Having established the basic 'scientific' hypotheses the next stage was to establish the framework of the service. Other issues, such as data exchange agreements and protocols were also established at this point. Our choice of service was determined by a number of factors amongst which were

- The potential for using Automatic Speech Recognition
- Usefulness in a multilingual environment
- Relevance to our parent Telecommunications companies
- Feasibility within the time frame of the project.

One service that met these all of these constraints was originally conceived as a pan-European multilingual help-line. This would provide help on how to use the telephone network in foreign countries. For example an Italian speaker visiting France would be able to obtain guidance, in Italian, on how to use the major facilities of the France Telecom network. Examples of the types of information to be provided were emergency and embassy numbers, country and area codes, national and international directory service numbers, country direct numbers, tariffs and how to use Credit and Telephone Cards.
Clearly such a service could be implemented with various degrees of sophistication ranging from a simple DTMF service to one using a full natural language interface. It was also apparent that each implementation would have to be 'tuned' to the services actually available in each host country.

By this stage the architectural structure of Fig 1 was beginning to emerge. However just how well this would work in all countries was not clear, so in parallel with addressing the first set of questions, the first experimental phase was designed to optimise the dialogue structure for the service environment of each.

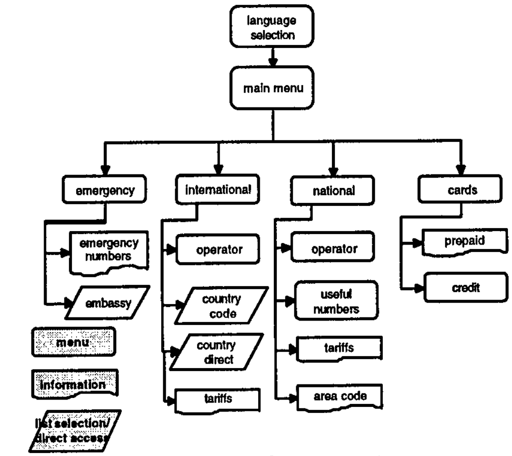


Fig 1 – Outline of the basic service structure

3. THE MONOLINGUAL EXPERIMENTS

3.1 General

The first series of experiments were mono-lingual experiments. In these all partners independently implemented and tested their platforms but adhered to the common structure and methodology.

The aim of these was to undertake experiments concerned with:

fixed/cellphone comparisons
prompt optimisation
recognition/Touch-tone comparisons
recognition threshold settings which are preferred

The following factors were used in each case
A total of 16 test conditions were devised by combining three recogniser settings + 1 'Touch-tone' with four dialogue structures. At least two different types of phone (GSM and fixed network) were then used in the test process.

3.2 Tasks

For each 'conversation' with the system subjects were given a task to complete. These were divided into two categories - those, which had answerable questions, and those which the system could not answer. The two types were chosen because in real systems many users request information which the system does not contain. We wanted to examine how different dialogue strategies coped with this situation and how users reacted to them . Examples of the answerable questions were:

What is the number for the French Embassy?
What number do you call for an operator?
What number do you dial for chargecard information?

The unanswerable questions were very similar – but the database did not support them e.g.:

What is the number of the Russian Embassy?
What number to you call to install a new phone line?
What is minimum charge for calls made using credit cards?

3.3 Test Procedures

In principle, with 16 test conditions/treatments it should be sufficient to undertake a fully balanced experiment with 16 subjects and 16 tasks. However the complicating factor of the answerable/unanswerable questions and the slight imbalance created by having 3 recogniser based system and one DTMF system meant that a modified design involving 16 subjects and 20 tasks were used.

This highlights the sequence of conditions that a typical subject experienced. Each subject was first given 3 'practice' sessions to ensure they were comfortable with the procedures. The above design is balanced in that every subject experiences all the conditions and all the tasks. However no subject experiences exactly the same combination of tasks and conditions as any other. The advantage of using such a balanced approach becomes evident when the analysis stage is reached for it becomes possible to partition the data in many different ways. For example exactly half of all calls will have been to fixed networks and exactly half to GSM networks. The balancing process guarantees that each of these groups will contain exactly the same set of tasks and exactly the same distributions of talkers.

How this helps in the analysis is shown below, but before that we examine the responses.

3.4 Responses

There are a number of well known subjective response scales available for this type of subjective procedure^2. However for our purposes none were immediately suitable. The added complication of task completion - which could be successful or unsuccessful had to be taken into account as there is an important distinction to be observed between satisfaction with the result obtained and satisfaction with the system used to obtain it.

The three dimensions identified were:

- Did the service deliver what it was supposed to?
- Was it easy to use?
- Was it pleasant to use?

To deal with this complication the following three simple subjective responses were collected after each transaction.

Please think first about the **quality of the result** you have just obtained, and mark one of the following to show your opinion.

- Fully satisfactory
- Satisfactory in the main, but left something to be desired.
- Unsatisfactory or misleading
- Irrelevant or positively wrong
- No result obtained at all

Now please think about your **satisfaction or dissatisfaction with the system** that you have just used to obtain this result.

How easy or difficult was this system to understand and use?

Allocate a figure of merit in the range 1 to 10 where 1 represents "Very difficult to understand or use" and 10 represents "Very easy to understand and use". [ ]

How pleasant or unpleasant was this system to use?

Allocate a figure of merit in the range 1 to 10 where 1 represents "Very unpleasant, slow or tedious to use" and 10 represents "Very pleasant and interesting to use" [ ].

Besides these subjective responses to each transaction, the following measurements were made:

Time taken per transaction.
Number of error-correcting dialogues entered.
Numbers of substitution or insertion errors.
Correctness or incorrectness of each result obtained.

Each task and result was presented separately to each subject who then had to respond on paper with each sheet collected after each call.

3.5 Analysis of results

The factorial design allows the results to be analysed in several ways. It also allows statistical tests to be undertaken to determine the significance of the components.

To illustrate the richness of the output data , some examples of these types of results are shown in Figures 3 and 4. These show the various objective (speed, error-rate and length) responses and the three subjective responses (Result, effort and Pleasant) for each case with the data partitioned between fixed and GSM results.

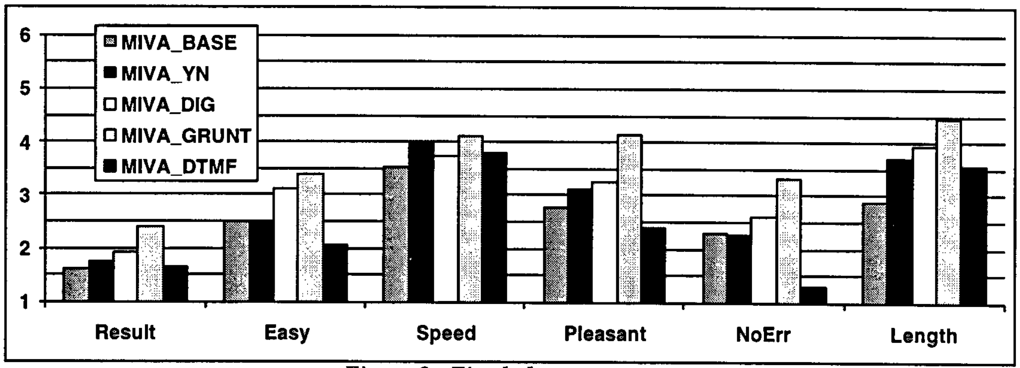


Figure 3 - Fixed phone responses

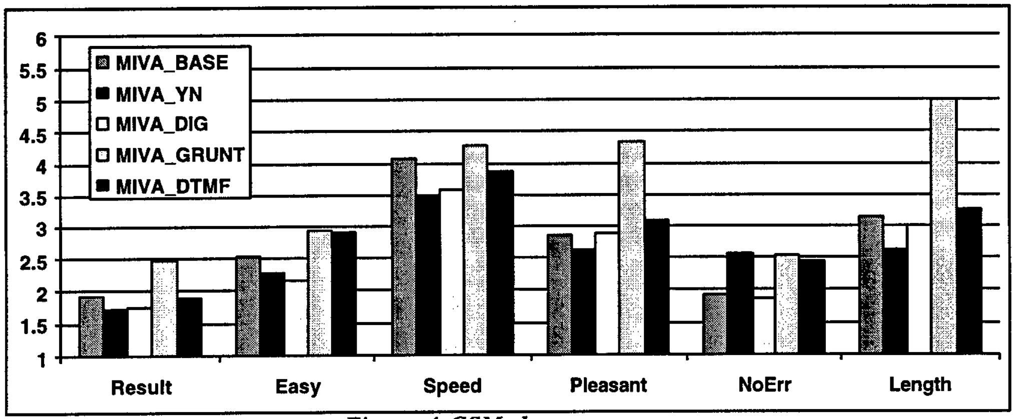


Figure 4 GSM phone responses

Direct observation shows that there are few substantial differences between the two sets of results and that the differences between, for example, dialogue types are much greater. Statistical tests were undertaken to establish any significant differences. The main conclusions from the analyses at this point were:

- All partners found that for menu driven systems of the type explored here, accessed over the fixed network, DTMF was easier to use than speech recognition.
- When the same services were accessed from mobile phones (mainly GSM) speech recognition was marginally preferred by all partners with the exception Deutsche Telekom where DTMF was still slightly preferred.
- 'Simple' recognition strategies e.g. using 'grunt' detection or numbered items are preferred less than proper word based recognition for all cases of fixed and mobile phone access.
- The absolute accuracy of recognition over mobile networks (as compared to the performance of the over the fixed network) was found to be slightly worse than that for fixed network in Deutsche Telekom and Italia Telecom. However no significant difference was found in the BT and Portugal Telecom systems.

4.MULTILINGUAL EXPERIMENTS

4.1 Introduction

The mono-lingual experiment had addressed the bulk of the technical questions but had not addressed any of the multi-lingual aspects. It had also tested out the individual platforms and allowed us to define the country specific dialogues.

The multi-lingual experiments built upon these results and answer the remaining questions by exploring

- Effects of cut-through
- What is the best way to prompt users to say a language
- Comparison of national versus International access
- Language/cultural differences.
- Predictive modelling of objective/subjective results.

As each partner had by now implemented and optimised their own local service in their own language the basic structures of the services were now well established. The next stage was to migrate to the totally multi-lingual environment. This meant that every suite of dialogue prompts in each language had to be translated, sent to the 'mother tongue' country for recording and then sent back. At the same time a data collection exercise to collect the necessary words to train the speech recognisers in every language had to be undertaken. There were, on average, about 70 words to be recorded per dialogue. Appropriate vocabularies had to be collected from a representative set of over 800 people in each country, labelled and stored in a form that would allow the specific service to be implemented in every location

Although the logistics of this seemed quite complex, in reality the use of digitised speech recordings stored as files on CD-ROMS proved so reliable that the recording and distribution processes ran extremely smoothly. To check quality at various stages the recognition components were tested in the laboratory with a common test set -country experiment.

With the five platforms each supporting the five language variants the experimental phase could start.

4.2 Multi-lingual Experimental design

Balanced experimental designs similar to those for the mono-lingual experiments were used as the basis for collecting objective performance data and subjective preferences.

This time the variables of interest were factored by the 5 languages and 2 dialogue types to give a total of 10

treatments. The dialogue types were either “Short and concise” or “Descriptive/verbose” and were further subdivided into technologies (by countries) to the extent that two classes of service could be identified. The first had essential information first (e.g. country direct number), followed by details on how to use it. This structure was implemented on the FT, BT and PT platforms.
The second had descriptive information first, followed by information content successively. This structure was typically implemented on the IT and DT platforms.
Also two countries (France and Italy) deployed 'cut-through' in their systems.

The total number of subjects recruited for the test was 100, a panel of 20 subjects per partner. At end of the experiment a total number of 900 calls had been collected and 886 questionnaires completed. Extensive objective data had also been collected.

4.3 Multilingual Analysis

As before, the results could be factored and analysed using analysis of variance techniques. For example comparisons could be drawn between the dialogue types, the use (or not) of cut through and any effects of language.

It was also possible to use the data to see if there were any 'learning' effects taking place. Evidence that there was is illustrated by the following analysis. Learning effects were tested by means of the Chi-square test between opinions for consecutive calls. Subjective measures were then compared once we had identified the number of calls a subject needed to place before he could be considered an expert. Figure 5 which is extracted from [3] illustrates the effect.

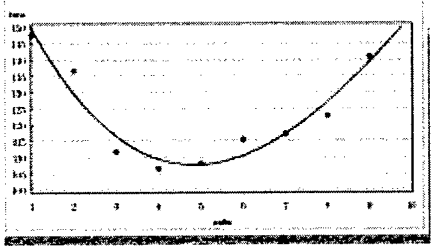


Figure 5. Transaction time as a function of order of call

4.4 Regression analysis.

Regression analysis is a technique used to determine the relationships between variables.
In this case we wanted to see if we could predict subjective responses from the various objectively measured parameters. From an initial examination of the data suggested that the subjective variables with the highest association with satisfaction were transaction time, number of utterances and number of correction turns. Using half of available data a series of regression analyses were performed to find an analytical model that best fitted the data for each parameter.

For the linear case functions and parameters found were:

Ease of Use = 4.440576 - 0.004778* transaction time
Learnability = 3.848315 - 0.080385*numbers of utterances
Pleasantness = 3.881013 - 0.003505*time
Effort = 3.914733 - 0.288068*corrections turn
Correctness = 0.576166 + 0.035409*word recognition rate;
Duration = 3.726034 - 0.005278* transaction time

4.5 Validation of the model

To validate the above the other half of the data was used. Below shows the results when the complementary data was applied to the ease of use data.

	N	Mean time	Observed	Prediction by linear regression
BT	111	96.51	3.89	3.98
DT	104	172.5	3.54	3.62
FT	116	68.0	4.47	4.12
IT	54	177.4	3.61	3.59
PT	101	116.6	3.71	3.88

The differences between observed and predicted values were tested by means of t-test. In both cases the test was not significant: the predicted mean value did not differ from the observed mean demonstrating external validity.

4.6 Observations

The regression methodology provided some further intriguing data concerning different language/nationality behaviours. For example English users strongly associated *satisfaction* with ‘transaction time’. Germans on the other hand associated *satisfaction* and *correct recognition feeling* with ‘transaction success’. Italian subjects associated *satisfaction, correct recognition* and *effort* ratings with ‘correction turns’ and ‘word recognition’. For the Portuguese subjects *learnability* was the only parameter associated with the objective ‘transaction success’. One conclusion may be that the questions were interpreted significantly differently by different groups.

5. CONCLUSIONS

We have described how standard methods of multi-factor experimental design have been adapted to evaluate the relative importance of various aspects of Interactive Voice response systems. The way in which subjective and objective data may be correlated and used as the basis of a predictive model has also been demonstrated. The model itself was then verified using a split data approach.

Although the project covered a great deal, there were some limitations. For example all experimentation was undertaken in a laboratory – as opposed to a market – environment. One exception to this was France Telecom who went one step further and undertook a public trial in the Musee de Lannion.

6. ACKNOWLEDGEMENTS

The authors gratefully acknowledge all the project participants that designed and implemented the MIVA system and collaborated for the realisation of the experiment: Sheyla Militello, *Joaquin Azevedo, Nuno Beires, Francis Charpentier, Mark Farrell, Eric Le Flour, Giorgio Micca, Karsten Schroede. We are especially indebted to Juan Siles (EURESCOM supervisor) for his valuable contribution to reviewing all the reports produced during the project.*

7. DISCLAIMER

This document may not reflect the technical position of all the EURESCOM Shareholders; its contents and specifications may be subject to further changes without prior notification. This document contains material, which is the copyright of some EURESCOM Project Participants and may not be reproduced or copied without permission. The commercial use of any information contained in this document may require a license from the proprietor of that information.

8. REFERENCES

[1] R.E. Kirk: “Experimental design procedures for the behavioural science”, Monterey, CA: Brooks-Cole Publishing.1982.

[2] Richards, D.L. “Telecommunications by speech” Butterworths 1973,.

[3] Militello S, Johnston R.D. "A Methodological study for the evaluation of a multilingual IVR user interface", Human Factors in Telecommunications, 17th Symposium, 1999.

A PLATFORM FOR MULTILINGUAL RESEARCH IN SPOKEN DIALOGUE SYSTEMS

Ronald A. Cole, Ben Serridge§, John-Paul Hosom‡, Andrew Cronk‡, and Ed Kaiser‡*

*Center for Spoken Language Understanding; University of Colorado – Boulder; Boulder, CO, 80309, USA
§Universidad de las Americas; 72820 Santa Catarina Martir; Puebla, Mexico
‡Center for Spoken Language Understanding (CSLU), Oregon Graduate Institute; Beaverton, OR, 97006, USA

*cole@cslu.colorado.edu; §serridge@alum.mit.edu; ‡{hosom,cronk,kaiser}@cse.ogi.edu

ABSTRACT

Multilingual speech technology research would be greatly facilitated by an integrated and comprehensive set of software tools that enable research and development of core language technologies and interactive language systems in any language. Such a multilingual platform has been one of our goals in developing the CSLU Toolkit. The Toolkit is composed of components that are essentially language-independent, and support research and development of recognition, understanding, text-to-speech synthesis, facial animation, and spoken dialogue systems. Portions of the Toolkit have already been ported to Italian, German, and Vietnamese. In addition, a complete Mexican-Spanish version of the Toolkit has been created, and is in daily use at the Universidad de las Americas in Puebla (UDLA). In this paper we outline some of the issues involved in porting the Toolkit to a new language, and describe why the Toolkit is well suited to multilingual adaptation.

1. INTRODUCTION

Speech communication occurs within social and cultural contexts, and is influenced by the perceptions, beliefs, attitudes, and backgrounds of the speakers. Research in spoken language systems requires participation by native speakers who understand not only the language, but also the subtle social conventions and cultural factors that enable natural communication. As a result, the best way to understand and model linguistically-related differences, create natural spoken-dialogue systems, and achieve acceptable machine translation between languages is through multinational collaborative research.

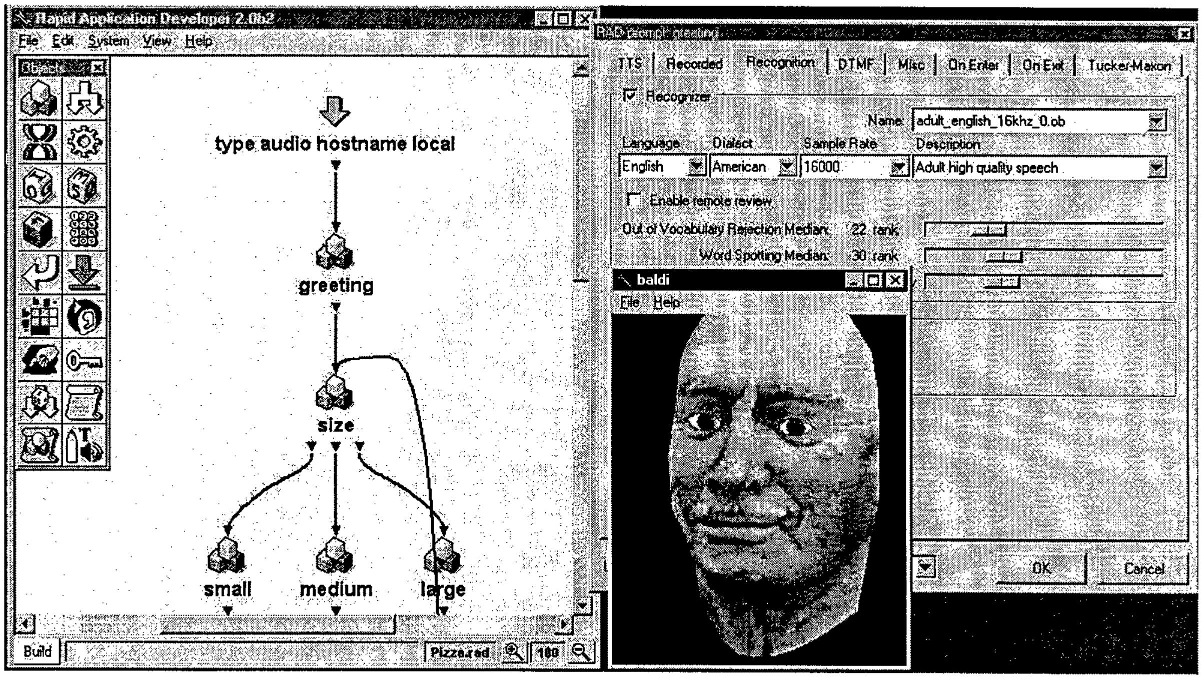
One of the main factors preventing more intensive multinational research is the "knowledge engineering bottleneck" — the massive costs associated with developing and deploying spoken language systems for each additional language and new application. These costs present formidable barriers to progress in human language technology.

Currently, spoken language systems development and research are limited to a few specialized laboratories because of the infrastructure and expertise required. The systems that are created in these laboratories are generally not portable; each new language and application requires collection of speech data, application-specific system development, and human engineering to create a graceful user interface. Data collection for training recognizers and for building language and dialogue models is costly and often must be done via "Wizard-of-Oz" simulation, with humans attempting to mimic the performance of a spoken language system. Such experiments are notoriously expensive and time-consuming. Consequently, all but the most fortunate students and researchers are denied the opportunity to explore this interesting frontier of human-computer interaction.

To break the knowledge engineering bottleneck and realize the potential of international, multilingual spoken-language research, it is necessary to develop available, usable, and powerful tools and corpora to engage and enable a generation of students to study, use, research, and develop language technologies and systems. These tools must be readily applicable to all languages of interest, so that there can be a common research framework. In general, most tools available today were designed by experts for use by other experts, and are not sufficiently tutorial to be used to train new researchers in undergraduate and graduate programs.

Our research efforts are aimed at overcoming these barriers, and the platform we use to integrate our advances is called the CSLU Toolkit. The Toolkit is freely available for research use from the CSLU Web site, and integrates speech recognition, natural language understanding, text-to-speech synthesis, facial animation, dialogue modeling, and spoken-language interface design in one package. The Toolkit is essentially language-independent; we have successfully ported the Toolkit to Mexican Spanish, and we are now developing a Brazilian Portuguese version of the Toolkit with colleagues at the Universidade Federal do Rio Grande do Sul, in Porto Allegre, Brazil.

The remainder of this article describes the CSLU Toolkit and its use a platform for multilingual research We hope that release of the Toolkit will remove entry barriers to research and education in human language technology, and enable researchers and students around the world to participate in creating the multilingual spoken-language systems of the future.



☑ Recognizer

Name: adult_english_16khz_0.ob

Language	Dialect	Sample Rate	Description
English	American	16000	Adult high quality speech

☐ Enable remote review

Out of Vocabulary Rejection Median: 22 rank

Word Spotting Median: 30 rank

Figure 1. Screen shot of the Rapid Application Developer (RAD), the animated face Baldi (with texture map), and a parameter window for setting properties of the recognizer.

2. THE CSLU TOOLKIT

The CSLU Toolkit is a comprehensive set of tools and technologies for learning about, researching and developing interactive language systems and their underlying technologies [1, 2, 3, 4]. It is available, free of charge, from the CSLU OGI Web site [5], along with CSLU's multi-language phonetically hand-labeled speech corpora. The Toolkit supports real-time interactive dialogues on standard off-the-shelf PC platforms running Windows (Solaris and Linux will be available soon). It provides a modular, open architecture supporting distributed, cross-platform, client/server-based networking. This flexible environment makes it possible to easily integrate new components and to develop scalable, portable speech-related applications.

The components of the Toolkit include both neural-network and HMM-based speech recognition systems, a natural-language semantic parser called PROFER, the Festival text-to-speech system, an anatomically accurate talking face called Baldi, and software for recording, displaying, labeling, and manipulating speech. The Toolkit also includes a GUI-based application developer called RAD and the documentation required to train HMM and neural-network based recognizers.

The tools are designed to enable inexperienced users to rapidly design, test and deploy spoken language systems. In addition to the pre-existing components, users can write their own C-level or script code for integration into the Toolkit. The recognition, synthesis, and natural language systems (and their tutorials) also support basic research and system development. Research advances can then be evaluated in real-world applications designed with the Toolkit's dialogue design tools.

Because the Toolkit is portable, runs on affordable off-the-shelf computing platforms, and provides both the knowledge (tutorials) and resources needed to conduct research, it removes some of the main entry barriers that currently prevent universities and research laboratories from establishing new programs in human language technology.

3. A PLATFORM FOR RESEARCH AND DEVELOPMENT OF MULTILINGUAL SPOKEN LANGUAGE SYSTEMS

The CSLU Toolkit is a proven platform enabling international collaboration in multilingual research and system development. Between 1996 and 1998, a joint NSF/CONACyT program supported collaboration between OGI and UDLA, the Universidad de las Americas, Puebla. The collaboration aimed to establish UDLA as a center of excellence in speech technology in Mexico, capable of educating students, of developing state-of-the-art Mexican Spanish spoken language systems, and of supporting research and education in language technologies throughout Mexico. These goals were accomplished.

As a result of our collaboration, the Tlatoa speech group has developed a complete Mexican-Spanish version of the Toolkit by collecting and transcribing corpora, implementing Mexican-Spanish recognition and text-to-speech systems, and converting Toolkit documentation to Spanish. UDLA now develops and distributes Mexican Spanish language resources, trains undergraduate and graduate students in language technology [6], publishes articles on speech research [7], and transfers knowledge and technology to other Mexican universities [8]. The collaboration has also produced industrial investment — a U.S. speech technology company has hired two UDLA students, has established a subsidiary in Puebla, and has made a substantial investment in the Tlatoa speech group (more than recovering the original investment from CONACyT).

The Toolkit has also been used successfully in European labs. Michael McTear has used the Toolkit to train undergraduate students at the University of Ulster to develop interactive language systems [9, 10]. Piero Cosi at the Istituto di Fonetica e Dialettologia Consiglio Nazionale delle Ricerche (Institute of Phonetics and Dialectology - National Research Council) has used the Toolkit to develop English and Italian speech recognition systems and compare Hidden Markov Model and neural network approaches [11, 12].

4. COMPONENTS OF THE TOOLKIT

In this section, the main components of the Toolkit are described. Issues in developing each component in a new language will be discussed, with examples from previous multilingual efforts where applicable.

4.1 Rapid Application Developer (RAD)

RAD is the Toolkit's high-level application developer. RAD's easy-to-use graphical authoring environment enables users to rapidly design and test spoken dialogue systems. It seamlessly integrates the core technologies of facial animation, speech recognition and understanding, and speech synthesis with other useful features such as word-spotting, barge-in, dialogue repair, telephone and microphone interfaces, and open-microphone capability.

RAD enables users to design interactive dialogues by specifying prompts, recognition vocabularies, and actions. Prompts can be either recorded or typed in as text, in which case they are produced as speech using the Toolkit's text-to-speech system. Both recorded and synthesized prompts are produced automatically by Baldi, the animated talking face [13]. Words or phrases to be recognized at any dialogue state are simply typed in by the system builder. Arbitrary actions can be associated with recognized utterances, such as producing a new prompt, displaying an image or retrieving and displaying information from a Web site. RAD contains many useful objects for retrieving, organizing and presenting information. In addition, users can develop new objects using the Tcl/Tk programming language. By connecting RAD objects, dialogues of arbitrary complexity can be designed. A sample screen shot of a RAD dialogue using Baldi with texture mapping is shown in Figure 1.

RAD currently includes both English and Mexican-Spanish recognizers and TTS voices. Addition of new recognizers and voices is easily done by creating new containers for the relevant objects (including the dictionary, if applicable) and storing them in the appropriate directories. Once these steps have been accomplished, RAD functions in the target language.

4.2 Facial animation

Baldi can be programmed within RAD to produce synthetic or recorded speech with different emotions. The face can be made transparent during speech production, revealing the movements of the teeth and tongue, and the orientation of the face can be changed while speaking to view it from different perspectives. Recently, a more complex and accurate tongue (consistent with electropalatography and imaging data), a hard palate, and three-dimensional teeth have been incorporated in Baldi. These features offer unique capabilities for language instruction — features that cannot be easily controlled in real faces.

Baldi is totally language-independent, in that he is controlled entirely by phoneme-level input. The input to Baldi consists of Worldbet [14] phonetic symbols, which are ASCII representations of the IPA and can represent all phonemes available in that alphabet. (The Worldbet system is used throughout the Toolkit, giving multi-language implementations a consistent phonetic representation.)

4.3 Speech Recognition

The Toolkit includes (a) English and Spanish digit and alpha-digit recognizers for recognizing sequences of digits and/or letters; (b) general-purpose English and Spanish recognizers for recognizing arbitrary words or phrases specified as text; and (c) a medium vocabulary English speech recognition system (MVCSR) that supports training of acoustic and language models for real time recognition of continuous speech with vocabularies up to 5000 words. The Toolkit supports research and education using several approaches to computer speech recognition, including artificial neural network (ANN) classifiers, hidden Markov models (HMM), and segmental systems. The Toolkit also includes step-by-step tutorials for training and testing new ANN and HMM recognizers.

The methods for training speech recognizers in the Toolkit are essentially language independent, with the selection of phonetic symbols and training corpora the only language-dependent parts. Pitch information is not currently used in the default feature set, but for tonal languages such as Mandarin or Vietnamese, the default feature set can be easily modified to include such information.

The Toolkit appears to be gaining acceptance as a platform for recognition research. In addition to English and Mexican Spanish recognizers, we are aware of

Toolkit recognizers developed for digit recognition in Italian, Vietnamese, and Korean. Consistent results have been observed across languages; for digit recognition, recognizers trained on telephone-band speech have word-accuracy levels of about 97% to 98%, and recognizers trained on microphone-quality speech have word-accuracy levels of about 99% [12, 15].

4.4 Natural Language Understanding

People do not always speak grammatically, and they often make false starts, or correct themselves as they are speaking. To parse this kind of spontaneous input requires a robust parser — that is, a parser that when confronted with such ill-formed input doesn't break, but finds the best allowable partial parse. Robust parsers, like Carnegie Mellon's Phoenix parser [16], are based on semantic *case-frame* architectures. They allow *slots* within a particular case-frame to be filled in any order, and allow out-of-grammar words to be skipped over. Thus, partial parses can be returned as *frames* in which only some of the slots have been filled. Typically, semantic case-frame parsers are implemented as chart parsers, and accept a transcript from the speech recognizer as their input. This requires separate grammars for the recognizer and the semantic parser, and limits the possibility of feedback between the two.

We have developed a semantic case-frame parser that runs as a finite-state machine rather than as a chart parser [17]. We believe this makes it more amenable to being tightly integrated into a speech recognizer, in such a way that the recognizer and semantic parser can share grammars and provide immediate feedback to each other. This tight integration is the aim of our current research. However, our initial version of Profer (which stands for Predictive, RObust, Finite-state parsER) can be used as a standard robust parser in a second-pass system, accepting the transcript produced by a recognizer. For example, using a grammar that defines sequences of numbers (each of which is less than ten thousand, greater than ninety-nine, and contains the word "hundred"), inputs like the following string of three numbers, which is rife with false starts and on-line corrections, can be robustly parsed by Profer [18]:

Input:

first I've got twenty ahhh thirty yaaaaaa thirty ohh wait no twenty twenty nine hundred two errr three ahhh four and then two hundred ninety uhhhhhh let me be sure here yaaaa ninety seven and last is five oh seven uhhh I mean six

Parse tree:

```
[fsType:number_type,
 hundred_fs:
  [decade:[twenty,nine],hundred,four],
 hundred_fs:
  [two,hundred,decade:[ninety,seven]],
 hundred_fs:
  [five,hundred,six]]
```

Profer is essentially a regular grammar parser. It allows the grammar writer to specify patterns in the input that should be "tagged" in the output parse tree as belonging to certain slots in a particular frame. The names of slots and frames are arbitrary — they can describe standard syntactic elements or task-specific semantic categories. Both tag-names and the patterns that define them are language independent. The grammar writer has free reign in this regard. Thus Profer is a language-independent tool, and has been used to define both English and Spanish grammars. A step-by-step tutorial has been developed for Profer to develop a conversational system for retrieving movie times and locations from a Web site.

4.5 Festival Speech Synthesis System

The Toolkit integrates the Festival text-to-speech synthesis system [19], a complete environment for learning, researching, developing, and using synthetic speech, including modules for normalizing text (e.g., dealing with abbreviations), transforming text into a sequence of phonetic segments with appropriate durations, assigning prosodic contours (pitch and amplitude) to utterances, and generating speech using either diphone or unit-selection concatenative synthesis. In addition, a graphical user interface enables users to "mark up" a text string to control many features of the resulting synthesized speech (e.g., rate, pitch, and amplitude) and to insert pauses, filled pauses, coughs, and sneezes.

During the summers of 1997 and 1998, researchers in the Speech Synthesis Research Group at OGI developed Spanish and German voices for use in the CSLU Toolkit. Students from the University of the Americas Puebla (UDLA), the University of Stuttgart, and the University of Bonn collaborated in these efforts. More information on these projects is available at http://cslu.cse.ogi.edu/tts.

While details vary, the overall process of developing a new voice is consistent between languages. As Festival is a concatenation-based synthesizer, a speech corpus must be designed and collected that optimally covers the target linguistic space. For example, a sample target linguistic space might be the phonemes of a language. In practice, such simple speech units are not used because they do not capture the coarticulatory effects between phonemes. Both the Spanish and German voices developed at OGI use the diphone as the basic unit of concatenation.

A promising technique known as unit selection is the focus of much ongoing research at OGI and other speech labs. In unit selection, longer — possibly non-uniform — "chunks" of speech may be extracted from a large, continuous-speech corpus. The goal of unit selection is to reduce the number of concatenation points in an utterance and increase the number of coarticulatory events captured in the speech in order to improve naturalness.

The process of developing text-to-speech corpora for waveform synthesis of new voices in new languages requires a series of steps. A protocol or script must be designed that contains at least one instance of each speech unit. Often the protocol is comprised of nonsense words or word pairs from which the diphones may be extracted. As not all phoneme-to-phoneme transitions exist in a given language, the advice of a native speaker or a trained linguist is exceedingly useful in keeping the protocol to a manageable size.

Once the protocol is optimally designed and the speaker selected, recording may proceed. High-quality recording is vital to the successful deployment of a new voice. The recording studio should be as anechoic as possible and possess high-quality microphones. A laryngograph is used to measure the impedance across the glottis during the session. These data are used to determine pitch marks, which are needed for smoothing concatenation points and altering prosody. The bulk of time invested in the development of the voices was spent separating and labeling these data.

For any language, rules must be developed to transform text into a sequence of tokens. For instance, the English text, "Dr. Suess spent $2.01 on Lorax Dr." may be represented by the tokens "doctor suess spent two dollars and one cent on lorax drive". Festival allows these rules to be easily scripted in Scheme, a dialect of the LISP programming language. In addition, mechanisms for determining the pronunciation of a token must be prepared. Since Spanish is a very consistent language, a set of letter-to-sound rules suffices. However, as the English and German languages are not particularly consistent, a lexicon must be found or created.

Finally, modules for the prediction of phoneme duration and pitch must be devised. These can be as simple as averages or they may be trained from data. Festival provides a number of tools for training prosodic modules from data.

Once all the above steps are complete, the voice may be defined within Festival. While this may be a challenging task for the first time Festival developer, once achieved for a particular language, the file formats and configuration files for each additional language are quite similar and readily created. In fact, the German synthesizer was speaking "guten tag" after only one day's work. The remaining month was spent collecting and preparing the speech data.

4.6 SpeechView

SpeechView is the Toolkit's interactive analysis and display tool. It allows users to create new waveform and label files, display data that are associated with a waveform (such as spectrograms or pitch contours), and modify existing waveforms and label files. It is used at CSLU for research, corpus development activities, and forms the basis for an interactive spectrogram reading class [20]. SpeechView supports simultaneous recording and subsequent annotation of auditory and visual speech data, and was recently used to collect bimodal speech data from over 250 children. SpeechView is entirely language-independent.

4.7 Perceptual Science Laboratory (PSL)

PSL provides a user-friendly research environment for designing and conducting multimodal experiments in speech perception, psycholinguistics, and memory. It enables users to manipulate auditory and visual stimuli; design interactive protocols for multi-media data presentation and multi-modal data capture; transcribe and analyze subjects' responses; perform statistical analyses; and summarize and display results. We plan to use PSL in our research to evaluate auditory visual synthesis for new languages. PSL, like SpeechView, is language-independent.

4.8 Programming environment

The Toolkit comes with complete programming environments for both C and Tcl, which incorporate a collection of software libraries and a set of API's. These libraries serve as basic building blocks for Toolkit programming. They are portable across platforms and provide the speech, language, networking, input, output, and data transport capabilities of the Toolkit.

5. CONCLUSION

The Toolkit has proven itself to be well suited for multilingual research in several areas. It is in use in over 300 laboratories worldwide, and has enabled research leading to over 200 publications.

In recognition, both English and Mexican-Spanish general-purpose recognizers have been created and are incorporated within the rapid application developer (RAD). Furthermore, the tutorial for training a digits recognizer has been used successfully by others in languages as diverse as Italian and Vietnamese. The semantic parsing tools in Profer are essentially language independent and are being used in both English and Mexican-Spanish applications.

In text-to-speech, we have developed Mexican-Spanish and German voices, the implementations of which were performed in one month, including the time required to collect and hand-label the diphone databases. In addition, we are currently refining a unit selection approach which is easily applicable to other languages and promises to improve naturalness.

Once the recognition and TTS components have been implemented in a given language, the graphical authoring tools enable rapid development of structured dialogue applications in that language. Finally, the components of the Toolkit can easily be interchanged, allowing quick substitution of an English recognizer with an Italian one, or German TTS with English. These factors all contribute to making the CSLU Toolkit powerful and easy to use in a multilingual environment.

6. ACKNOWLEDGEMENTS

This work was made possible by NSF Challenge award NSF/CDA-9726363, NSF CARE award UCB/NSF/153-0584, NSF grant NSF/IIS-9614217, ONR/DARPA grant ONR/N00014-94-1-1154, NSF Career award IIS-9875950, the support of Intel Corporation, and the CSLU Member companies.

REFERENCES

[1] S. Sutton, D. G. Novick, R. A. Cole, and M. Fanty. Building 10,000 spoken-dialogue systems. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Philadelphia, PA, October 1996.

[2] S. Sutton, R. Cole, J. de Villiers, J Schalkwyk, P. Vermeulen, M Macon, Y Yan, E. Kaiser, B. Rundle, K Shobaki, P. Hosom, A. Kain, J Wouters, M Massaro, and M Cohen. "Universal Speech Tools: the CSLU Toolkit." In Proceedings of the International Conference on Spoken Language Processing (ICSLP), pages 3221-3224, Sydney, Australia, November 1998.

[3] R. Cole, S. Sutton, Y. Yan, P. Vermeulen, and M. Fanty. Accessible technology for interactive systems: A new approach to spoken language research. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Seattle, Washington, May 1998.

[4] R. Cole. "Tools for research and education in speech science." In Proceedings of the International Conference of Phonetic Sciences, San Francisco, CA, Aug 1999.

[5] http://cslu.cse.ogi.edu/toolkit

[6] A. Barbosa, "A new Mexican Spanish voice for the Festival text to speech system." Masters Thesis, May 1997, UDLA-Puebla.

[7] B. Serridge, R. Cole, A. Barbosa, A. Vargas, and N. Munive. "Creating a Mexican Spanish Version of the CSLU Toolkit" Proceedings of the International Conference in Spoken Language Processing, Sydney, Australia, November 1998.

[8] B. Serridge, "An Undergraduate Course on Speech Recognition Based on the CSLU Toolkit." Proceedings of the International Conference in Spoken Language Processing, Sydney, Australia, November 1998.

[9] M. McTear, "Modelling spoken dialogues with state transition diagrams: experiences with the CSLU toolkit." Proc 5th International Conference on Spoken Language Processing, Dec 1998, Sydney, Australia, 1223-1226.

[10] M. McTear, "Using the CSLU toolkit for practicals in spoken dialogue technology." M.A.T.I.S.S.E. workshop on Method and Tool Innovations for Speech Science Education, April 1999, University College London, April 16 – 17.

[11] P. Cosi, J.P. Hosom, J. Schalkwyk, S. Sutton, and R. A. Cole, "Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-Based Recognizers", 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications (IVTTA-ETWR98), Turin, Sep. 1998, pp. 135-140.

[12] P. Cosi and J.P. Hosom "HMM/Neural Network-Based System for Italian Continuous Digit Recognition", In Proceedings of the International Conference of Phonetic Sciences (ICPhS), San Francisco, CA, August 1999.

[13] D. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*, MIT Press, Cambridge, 1998.

[14] J. Hieronymus, *ASCII phonetic symbols for the world's languages: Worldbet.* AT&T Bell Laboratories, Technical Memo, 1994.

[15] J.P. Hosom, R.A. Cole, P. Cosi "Improvements in Neural-Network Training and Search Techniques for Continuous Digit Recognition", Australian Journal of Intelligent Information Processing Systems, accepted for publication.

[16] W.H. Ward, "The Phoenix System: Understanding Spontaneous Speech", IEEE ICASSP, April 1991.

[17] E. Kaiser, M. Johnston, and P. Heeman, "Profer: Predictive, Robust Finite-State Parsing for Spoken Language." In Proceedings of ICASSP, Phoenix, Arizona, March 1999.

[18] E. Kaiser, "Robust, Finite-State Parsing for Spoken Language Understanding", in the 37th Annual Meeting of the Association for Computational Linguistics (ACL99), June 1999.

[19] A. W. Black, P. Taylor, and R. Caley, "The Festival Speech Synthesis System," http://www.cstr.ed.ac.uk/projects/festival.html, 1998.

[20] T. Carmell, J.P. Hosom, and R. Cole. "A computer-based course in spectrogram reading." In Proceedings of ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education, London, UK, Apr 1999.

AUDITORY FEATURES UNDERLYING CROSS-LANGUAGE HUMAN CAPABILITIES IN STOP CONSONANT DISCRIMINATION

Eduardo Sá Marta, Luis Vieira de Sá
email: EMARTA@CO.IT.PT
Dep. Engenharia Electrotécnica, FCTUC (Universidade de Coimbra)
Instituto de Telecommunicações - 3030 Pólo de Coimbra, Portugal

ABSTRACT

For some phonemic distinctions human listeners exhibit a marked cross-language capability, in that they are capable of highly correct classification in relation to sounds (like CVs or VCVs) uttered by speakers of another language. This is particularly true regarding distinctions that are perceived in a more categorical fashion, like that of 3-way PLACE discrimination in stop consonants. It is plausible that the reason for this is a mostly common (across languages) auditory basis for human communication of this discrimination. Also, human communication of this discrimination is notably impervious to non-drastic variations in the frequency-transfer curve, which suggests that the relevant auditory features must have some inherent insensitivity to these variations.

Models for two specialized auditory cells (onset cells with wide receptive fields, which can detect weak onsets synchronized across frequency, and sequence cells which detect frequency-ascending sequences composed of two onsets) were refined for the discrimination of DENTAL vs LABIAL stop consonants and applied to large spelling databases in Portuguese, German, and U.S. English. Similar discriminatory capability was observed both for German and U.S. English. Integration with a 3rd auditory feature resulted in error scores of approximately 2% when exactly the same model is applied to either German or U.S. English sounds.

1 - INTRODUCTION

1.1 – Human cross-language capabilities in stop consonant PLACE discrimination

It is well established that stop consonant PLACE discrimination is very well carried across languages (contrary to the voiced/unvoiced distinction, which for instance carries very poorly from U.S. English speakers to Portuguese listeners).

In a recent study [1], it is shown that native Korean listeners are capable of discriminating stop consonant PLACE, as uttered by U.S. English speakers, with less than 1% errors. This result, however, was obtained with utterances previously selected to be consistently classified, as well as to be the highest rated in goodness judgements, by native listeners of U.S. English. Thus, the results that might be obtained with an unselected mix of speakers, comprising speakers of good to below average intelligibility, could be somewhat poorer. That is, speakers of less good intelligibility strain the classification capabilities of native listeners, but these listeners are still able to maintain a very high score of correct recognition. Non-native listeners, on the other hand, may incur in significant error rates when faced with these poorer speakers.

The above discussion is useful in that it suggests expectations for a wholly correct model of human recognition (for listeners of a particular language, and for the above mentioned task): error scores as low as 1% may not be attainable, when the model is faced with databases of a different language which include a significant *proportion of speakers of less good intelligibility*. On the other hand, if this proportion is not high (say, less than 5%), the error rate should not be much higher than 1% (say, on the order of 2%) and should not suffer appreciably from mild variations of the frequency-transfer curve (say, on the order of ±2dB/octave in the range above 1KHz).

1.2 – General assumptions about human phoneme communication

These assumptions have been given elsewhere [2] but are recast here along with some additional considerations. We are bearing in mind communication tasks – such as spelling, and communication of nonsense words – in which humans exhibit a clear capability of speaker-independent phoneme communication. Nonetheless, the mechanisms crucial to this capability will obviously also be operative in word or sentence communication - though they may then be used for the communication of other speech units.

In the former tasks, there emerges – with very clear contours - the paradox of constancy of perception, in spite of variation of form. That is, the same CV, uttered by different speakers, presents very diverse acoustic forms (so diverse that extensively trained automatic recognizers incur – persistently - significant error rates) whereas human listeners correctly recognize the consonant, with apparent ease. The paradox is heightened when we consider that recording sounds through different microphones, or including speakers native of a different language (provided these are of "good quality") does not diminish appreciably the performance of human listeners, while wreaking havoc with automatic recognizers' performance.

To solve the paradox, we propose to consider the following visual communication analogy:

Suppose that a person is asked to draw pictures of a small set of fruits (pineapple, banana, orange, ...) just good enough to be correctly recognized when briefly flashed on a screen.

One particular drawer might present the PINEAPPLE texture very markedly; this will allow him to relax, for instance, the contour of the pineapple... ...which may even be rendered in a form ambiguous between PINEAPPLE and ORANGE Another drawer might "synthesize" a weakly marked texture, but then trace the contour in a very marked way.

In this "thought experiment", it may also be expected that *to draw a well perceived PINEAPPLE, a drawer may produce a texture that is much more marked than in any real pineapple (thus getting away from any conceivable category centroid), and by that he will still be aiding correct recognition*

This analogy suggests that for each phonemic distinction there are multiple *information carriers (ICs)* - or *cues*, or *features* - evaluated independently of each other, all being orthogonal to between-categories boundaries and that there exist, among the cues, trade-off relations that may extend to the point of alternativity. It is even conceivable that two different speakers may successfully communicate the same CV using entirely disjoint *ICs* ; this might be the case if a new speaker undergoing speech acquisition finds especially easy to emit a particular *IC* with high "intensity": this speaker may then "rest satisfied" and relax the emission of other *ICs* to the intended category. This plausible process is reminiscent of natural selection [3]: the well-known case of the panda's thumb, which achieves functional success (grasping action) with no morphological conformity (no real thumb) is particularly enlightening with respect to the paradox.

Another concept from natural selection which may be relevant to the phoneme communication problem is that of *exaption*, that is, the "seizing" by a new function (phoneme communication) of biological mechanisms that evolved previously as adaptations to other tasks (such as recognizing species calls in some distant animal ancestor, or as an even more basic survival-enhancing acoustic detection ability). This makes it likely that some of the *ICs* are mostly direct expressions of the acoustical metrics computed by some "hard-wired" (that is, not substantially modified in response to speech use) neural assemblies.

Use of several *ICs* pointing to the same category achieves redundancy and robustness to signal degradation: when degradation is not drastic, some *ICs* may be obliterated, but if some others survive, correct recognition by listeners will still be obtained.

The speakers also want to accommodate articulatory ease, indulge articulatory variability induced by various motivations (the conveyance of a personal speaking style, emotional status, etc.) – all of this is made possible by the extensive trade-offs between the several *ICs* for the same category.

1.3 – The set of *Information Carriers* (*ICs*) for human communication of the PLACE distinction in stop consonants

Characterization of this set is the subject of our on-going research, but the following *ICs* are thought to be well stabilized; further additions may have to be made, but their importance will be of a secondary degree.

Introduction of the *ICs* was driven by the need to explain the perception of PLACE in natural or edited sounds when no explanation could be found in terms of the *IC*s known at a particular time in the research undertaking. It order to provide a substantial number of such "driving sounds", the need for considering several languages was recognized early on; for a single language, most speakers conform to acoustic regularities particular to that language and the number of sounds that provide a clear-cut challenge for explanation of their perception is very limited.

The current characterization of each *IC* is the result of an hypothization endeavor, followed by satisfactory results in the application of a model of the *IC* to a large number of sounds.

Acceptable *ICs* must be biologically plausible and must exhibit some degree of independence to non-drastic variations in the frequency-transfer curve. Some *ICs* may correspond closely to metrics computed by some specialized auditory cells – in this case, the neural algorithms computed by these cells may yield a high selectivity in frequency and/or in time. Other *ICs* may correspond to *speech schemas* [4] and thus must be expressible in terms of plausibly auditory-salient representations such as gross integration of energy.

The (current) set of *ICs* for PLACE discrimination into the three categories LABIAL, DENTAL and GLOTAL/VELAR is then:

LABIAL-IC1: *ascending sequence in the F2/F3 zone*. This is assumed to be evaluated by ascending sequence cells such as those that have been found in the primary auditory cortex of primates [5]. Since the abruptness of onset is the most important characteristic of each of the two components of the sequence, insensitivity to non-drastic variations of the frequency-transfer curve is assured. There are many references in the perception literature to an "ascending" quality being a cue for LABIAL (see for instance [6]).

LABIAL-IC2: *ascending trajectory of the dominant low-frequency skirt in the F2-F3 zone*. For this *IC* there is also a two-times comparison but the "after" term is to be evaluated through temporally-gross integration, and the onset of the vowel functions as a temporal marker signaling this "after" term.

LABIAL-IC3: *complete or near-complete absence of unvoiced energy prior to vowel onset*. This is similar to the "burstless" quality referred in [7] as a cue for LABIAL. There are a number of studies in the literature that also concur in this finding, many of which are also cited in [7] . However, we added the detail that high-frequency energy occurring just at the vowel onset may provide a non-burstless percept. This *IC* is thought to be evaluated through temporally-gross integration.

LABIAL-IC4: *initial brief (<3-5ms) "vertical bar" in the spectrogram, followed by no significant high-frequency energy*. This is thought to be evaluated with the help of wide-receptive field onset cells, in conjunction with temporally-gross integration of energy.

DENTAL-IC1: *initial tone-burst-like segments (>6-8ms) (usually corresponding to the initial aspirated or voiced segments of F2 or F3) of very thin bandwidth.* This *IC* is primarily based on the output of hypothetical cells similar to *level-tolerant* neurons [8], although there seems to be also a temporal windowing mechanism involving onset cells.

DENTAL-IC2: existence of upward inflections in the spectrum, occurring at "high" (>3.5KHz) frequencies during the *burst+aspiration* segment. This is – albeit distantly - related with [9]. The metric for this *IC* is assumed to be dependent on auditory cells exhibiting marked lateral inhibition from the lower sideband.

DENTAL-IC3: segment prior to vowel release (that is, the *burst+aspiration* segment) having a considerably stronger high-frequency content then the ensuing vowel. This is likely to depend on temporally-gross energy integration, and on the use of the vowel onset as a temporal marker to distinguish the 2 terms of the comparison.

DENTAL-IC4: grossly equifrequencial sequence in the F2/F3 zone. Assumed to be based also on sequence cells of the primary auditory cortex.

GLOTTAL-IC1: descending sequence in the F2/F3 zone. Also based also on sequence cells of the primary auditory cortex

GLOTTAL-IC2: Strong onset of "compact energy" in the F2-F4 zone, followed briefly (<10-20ms) by abrupt offset. No specific auditory cells have been found in the literature to account for the evaluation of such a metric, but their existence has some biological plausibility.

GLOTTAL-IC3: descending trajectory of the dominant low-frequency skirt in the F2-F3 zone. The fact that such a trajectory will continually meet unadapted cells in the auditory nerve provides a basis for its auditory evaluation.

The above characterizations, and the present state of development of models for some of the *ICs* has been the result of extensive studies with natural and edited sounds. As a first step, we tried to predict the perception of sounds (of unknown PLACE) based on inspection of several spectral displays, and the estimation of how the relevant auditory structures would react to the sound. This in turn led to the development of fuzzy-logical, auditorily-plausible, models for some of the *ICs*.

We develop/refine the models through inspection of their results in 5 sets of sounds: an in-house research database of /ti/ and /pi/ sounds from 33 Portuguese speakers (representative of Portuguese unvoiced stops), the letters "T" and "P" from the first set (30 speakers, 120 sounds) of the Oregon Graduate Institute ISOLET Database (representative of U.S. English unvoiced stops), the letters "D" and "B" from the first set of ISOLET (representative of U.S. English voiced stops), the letters "T" and "P" from the first 50 speakers of the Bavarian Archive for Speech Signals PHONDATA1 Database (representative of German unvoiced stops), and the letters "D" and "B" from the same 50 speakers (representative of German voiced stops). It is to be emphasized that for each *IC* the same model is used throughout, with no adaptation whatsoever, and that the different sets have obviously used different microphones, as well as recording conditions.

2 – AN *INFORMATION CARRIER* FOR THE LABIAL CATEGORY, BASED ON ONSET CELLS

In this section, a model for **LABIAL-IC4** is discussed, along with its motivation.

In our research towards being able to predict the perception of sounds of unknown PLACE, we came across some sounds ("P" and "B" in spelling databases in German and U.S. English) whose LABIAL perception seemed more robust than could be explained in terms of the other three *ICs* for LABIAL (which were uncovered, and characterized, first). More definite conclusions could be extracted from some particular sounds which lent themselves to filtering or editing operations that clearly removed (or greatly diminished) the other *ICs* for LABIAL; many of these sounds maintained a clear LABIAL perception, raising the need for another LABIAL *IC*.

The common acoustical trait among these sounds was the presence of an initial "vertical bar" in the spectrogram followed by (at least) a few milliseconds with little energy across higher frequencies. One difficulty in the way of making this observation was that most often the "vertical bar" seemed to be of such low energy (relative to the rest of the speech signal) that at first it seemed improbable that it would play a significant part in perception.

But it was realized that some onset cells in the cochlear nucleus could exhibit an extremely wide receptive field, measured using the concept of *two-tone facilitation* [10] and that this could result in measurable responses even with the "best-frequency" tone being as low as 30dB below threshold. So, if it turns out that a fair proportion of LABIAL stops are capable of exciting these cells, while non-LABIAL stops are not, it is clearly conceivable that this came (during the evolution of languages) to constitute a valuable *IC* for LABIAL.

Since there are varied types of onset cells (with differently wide receptive fields), and members of each type may be found with central frequencies all along the audible range, there remains the question of establishing the characteristics of those onset cells that are mobilized for LABIAL-IC4. Cells sensitive to very low frequencies (say, below 2KHz) would tend to give unreliable information, since acoustical accidents due to non-speech noise are apt to cause excitation of such cells; we considered only cells with receptive fields extending upwards from 3.5KHz, up to 7.0KHz. Another issue is the frequencial width of the receptive field; we considered a fixed width of 1400Hz (a point which is to be refined in the future).

The essence of the neural algorithm for onset cells is the summation of the outputs from a large number of auditory nerve cells (spanning a wide frequency range), occurring simultaneously. It is possible that for some cells contributions emanating from a restricted frequency range will not suffice to excite the cell, however strong these contributions (it is even possible that a very strong frequency-local contribution will turn off the cell, through the hypothetical mechanism of *shunting inhibition*).

We implemented a fuzzy-logical model to account for these dependencies. For direct comparison with commercial spectrographic displays and sound editing software, we use simple FFT spectra as the input representation. The speech signal is represented by FFT spectra calculated, with a Hamming window, over frames of 11.6ms, with a 3-ms frame advance. Thus the input matrix is composed of points E(F,T) where F=fx86Hz and T=tx3ms. At each such point, we computed the ***Unadapted-Increment***(F,T) considering the energy at point (F,T) and energy previously occurring at frequencies proximal to F. ***Synch-Increment***(F,T) was computed with a metric similar to summation applied to ***Unadapted-Increment***(F',T) with F' spanning from F to F+1400Hz. In this summation-like metric, the contribution of outstanding peaks is subject to limitations. The most adequate form of these limitations is still being studied; for instance, the intriguing possibility that an extremely outstanding peak might actually decrease the response of the cell, through *shunting inhibition*, is for the time being kept open.

It is interesting to note that this metric is unavailable to conventional automatic recognizers, since their input representation has, as a rule, much poorer time resolution than used here.

The model was refined (in the version reported here, only about 10 parameters were explored) primarily using U.S. English "P" and "T" sounds and was applied unaltered to German and Portuguese "P" and "T" sounds. The histograms for U.S. English are presented below:

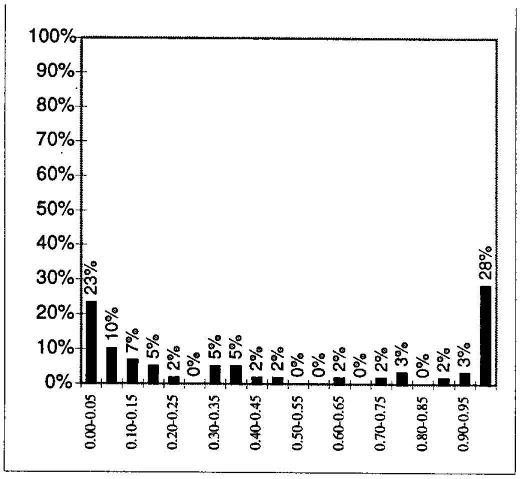


Figure 1 – Histogram for the fuzzy variable expressing LABIAL IC 4 for 60 U.S. English "P" sounds (Isolet, 1st set)

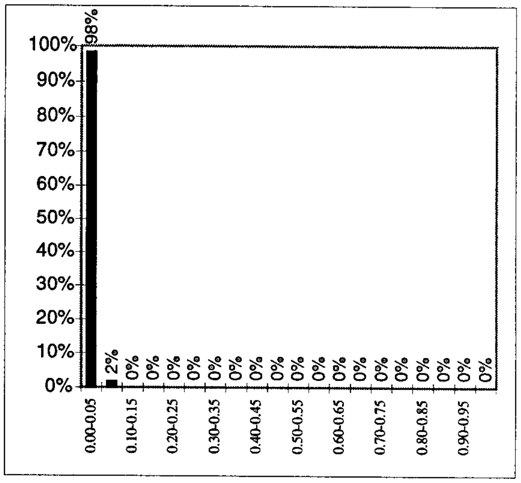


Figure 2 – Histogram for the fuzzy variable expressing LABIAL IC 4 for 60 U.S. English "T" sounds (Isolet, 1st set)

From these histograms, it is apparent that significant to high values in this metric only occur for LABIAL sounds, and not at all for DENTAL sounds, making it an obviously useful *information carrier* for the discrimination between these two categories. It is evident that the metric exhibits a generous "exclusively LABIAL" range of medium to high values, which range is only attained by LABIAL sounds.

The results obtained applying the same model to German sounds are given below:

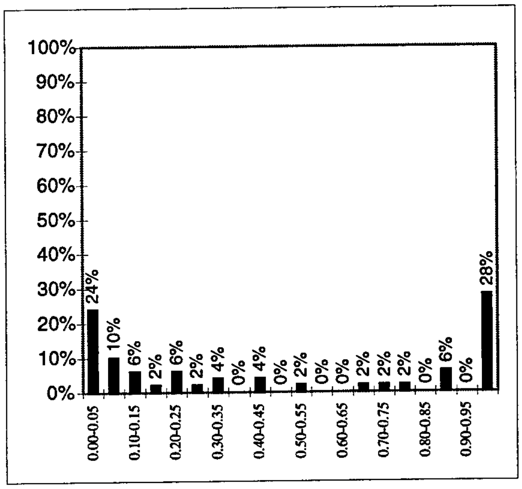


Figure 3 – Histogram for the fuzzy variable expressing LABIAL IC 4 for 50 German "P" sounds (PhonData1, first 50 speakers)

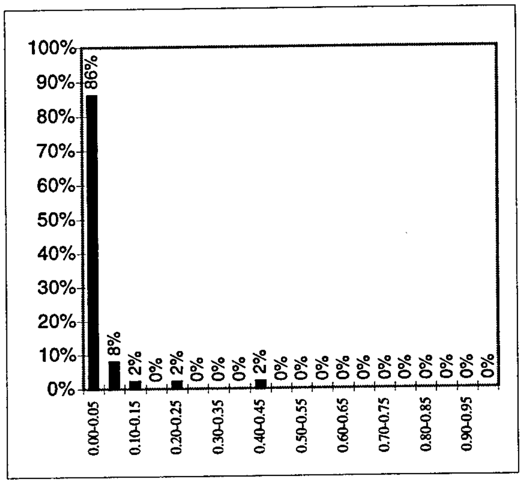


Figure 4 – Histogram for the fuzzy variable expressing LABIAL IC 4 for 50 German "T" sounds (PhonData1, first 50 speakers))

The results for German are similar to those for U.S.English. The "exclusively LABIAL" range is here somewhat spoiled by a single sound with a mark at 0.42 (sound "hdbdT" – PhonData1 labels) but that likely is the result of imperfect refinement of the model.

The results for Portuguese, however, are very poor: "P" sounds almost never elicit significant values in the metric. But this is not surprising, since "P" sounds in Portuguese have very weak and short *burst+aspiration* segments; in fact, it is uncommon in Portuguese for these segments exceeding 15ms in duration, whereas for U.S. English durations in excess of 100ms are frequent.

3 - INTEGRATION OF DIFFERENT *INFORMATION CARRIERS*

Even granting success in modeling the different *ICs* , the problem of their integration adds another layer of complexity. We will simply show – using the simplest possible fuzzy-union operator (the *maximum*) - how two different LABIAL *ICs* yield discrimination superior to that of the better of those *ICs* .

The LABIAL *IC* which has the better discriminatory power is LABIAL-IC1: *ascending sequence*. Histograms for Isolet 1 "P" and "T" are shown below:

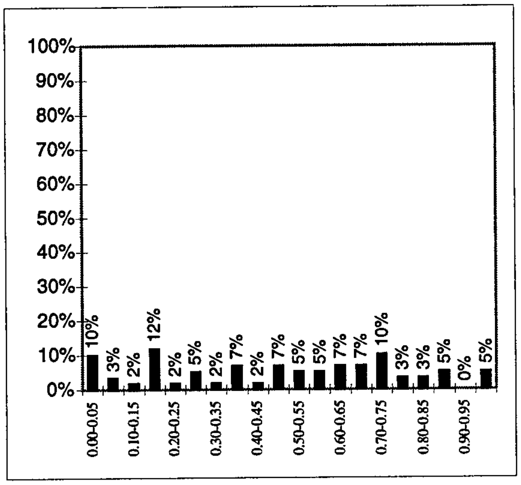


Figure 5 – Histogram for the fuzzy variable expressing LABIAL IC 1 for 60 U.S. English "P" sounds (Isolet, 1st set)

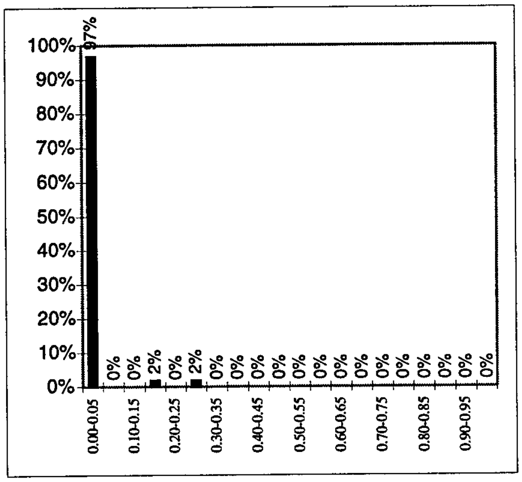


Figure 6 – Histogram for the fuzzy variable expressing LABIAL IC 1 for 60 U.S. English "T" sounds (Isolet, 1st set)

Simply taking the maximum of the two fuzzy variables expressing LABIAL IC 1 and LABIAL IC 4 yields the following histograms:

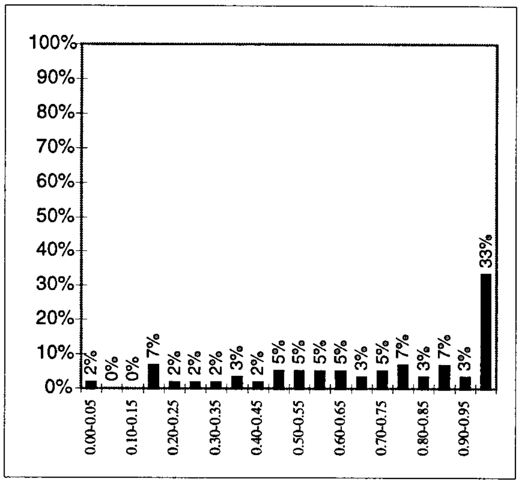


Figure 7 – Histogram for the maximum of LABIAL IC's 1 and 4 for 60 U.S. English "P" sounds (Isolet, 1st set)

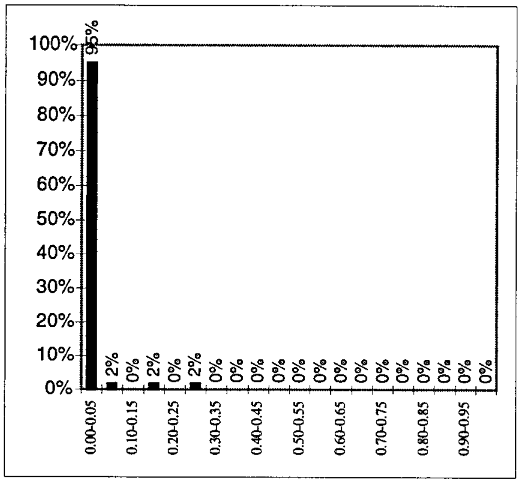


Figure 8 – Histogram for the maximum of LABIAL IC's 1 and 4 for 60 U.S. English "T" sounds (Isolet, 1st set)

Bringing in more *ICs* further improves discrimination. Performing fuzzy intersection of the above maximum (of LABIAL IC1 and LABIAL IC4) with the fuzzy variable expressing DENTAL IC-3 results in a fuzzy variable which, thresholded at 0.15 yields 1.7% "P" vs. "T" discrimination errors for U.S. English and 2% for German.

4 - CONCLUSIONS

A small number of *Information Carriers*, each with a reasonably simple characterization in terms of known auditory processes, is shown to be able to approach human capabilities in the cross-language communication of stop consonant PLACE. This was shown through modeling of some of the *Information Carriers* relevant to the LABIAL vs. DENTAL distinction.

Low error scores were maintained not only across languages, but also in spite of differences in recording settings that exist between databases. This suggests that the proposed metrics are substantially insensitive to non-drastic variations in the frequency-transfer curve.

Further work is going on in connection to *Information Carriers* relevant to the discrimination of GLOTAL/VELAR PLACE.

ACKNOWLEDGEMENTS

The research reported in this paper has been conducted under the Research and Development Contract Praxis 2/2.1/TIT/1558/95 of the PRAXIS XXI Program of the Junta Nacional de Investigação Científica e Tecnológica.

REFERENCES

[1] - Anna Marie Schmidt, "Cross-language identification of consonants. Pat 1. Korean perception of English", J. Acoust. Soc. Am. 99, pp.3201-3211, 1996

[2] - Eduardo Sá Marta, Luis Vieira de Sá - "Auditory cells with frequency resolution sharper than critical bands play a role in stop consonant perception: evidence from cross-language recognition experiments" - Proceedings of the *NATO Advanced Study Institute on Computational Hearing,* Il Ciocco, Italy, 1998

[3] – Gary Cziko, "Chapter 11 – The Evolution, Acquisition, and Use of Language" in "Universal Selection Theory and the Second Darwinian Revolution", MIT Press, 1995

[4] – Bregman, A.S., "Auditory scene analysis: the perceptual organization of sound", MIT Press, 1990

[5] - H. Riquimaroux, "Processing of sound sequence in the auditory cortex" - Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, Keele University (UK) - 1996

[6] A. Lahiri et al - A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: evidence from a cross-language study - J. Acoust. Soc.Am. 76, pp.391-404, 1984

[7] Smits,R., Bosch,L., Collier,R., "Evaluation of various sets of acoustic cues for the perception of prevocalic stop consonants. I. Perception experiment" J.Acoust. Soc. Am. 100 (6), pp.3852-3864, 1996

[8] Suga,N., Zhang,Y. and Yan,J., "Sharpening of Frequency Tuning by Inhibition in the Thalamic Auditory Nucleus of the Mustached Bat", Journal of Neurophysiology,Vol.77, pp.2098-2114, 1997

[9] - Stevens, K.N., and Blumstein, S. E. - "Invariant cues for place of articulation in stop consonants", J. Acoust. Soc. Am. 64, 1978, 1358-1368

[10] - Dan Jiang, Alan R. Palmer, Ian Winter – "Frequency extent of two-tone facilitation in onset units in the ventral cochlear nucleus" – Journal of Neurophysiology, vol.75, pp.380-395, 1996

USES OF THE DIAGNOSTIC RHYME TEST (ENGLISH VERSION) FOR PREDICTINGTHE EFFECTS OF COMMUNICATORS' LINGUISTIC BACKGROUNDS ON VOICE COMMUNICATIONS IN ENGLISH: AN EXPLORATORY STUDY

William D. Voiers
Dynastat, Inc.
2704 Rio Grande, Austin, TX 78705 USA
bvoiers@aol.com

ABSTRACT

Recordings of Diagnostic Rhyme Test (DRT) materials by native talkers of English (American), German and French were presented under undegraded and degraded conditions to English speaking listening crews of three national origins: American, German and French. The results were analyzed for the effects of the talker's native language, the listener's native language and all permutations of the two on scores yielded by the DRT. With undegraded speech, the total number of errors was lowest when the talkers were American, regardless of the nationality of the listeners, and when the listeners were American, regardless of the nationality of the talkers. On average, French talkers yielded the lowest DRT scores, but the interaction of talker nationality and listener nationality was significant. Errors of discrimination with respect to *voicing, sustention, sibilation* and *graveness* occurred most often.

Keywords: Intelligibility, Diagnostic Rhyme Test, multi-lingual interoperability

1. INTRODUCTION

Many factors potentially contribute to errors in speech communication in circumstances where the communi-cators are required to communicate in other than their native languages, as is frequently the case in civilian and military aviation communications. These factors include language differences in syntactical and grammatical rules. They also include differences in the phonemic alphabets of the various languages involved. Comparisons of the phonemic alphabets of the languages involved may permit identification of some of the more important sources of mis-communication, i.e., speech elements not common to the native languages of the communicators involved. Such comparisons do not, however, permit quantitative predictions regarding communication failures, nor do they permit distinctions between communication failures due to errors of articulation and those due to errors of perception — distinctions between failures due to the talker and those due to the listener.

2. PURPOSES

The purposes of this study were (1) to demonstrate the sensitivity of the Diagnostic Rhyme Test [1, 2] to the effects of communicator differences in linguistic background on voice communications conducted in English, (2) to evaluate the relative contributions of the talker's and the listener's linguistic backgrounds to voice communication failures and (3) to identify the speech elements and/or features most susceptible to misarticulation or misperception by non-native talkers of English.

3. METHODS AND MATERIALS

3.1 Speech materials

The speech materials used for this study were recordings of the test words of the Diagnostic Rhyme Test (DRT-IV). Although originally designed to aid communication scientists and engineers in pinpointing specific system defects or malfunctions, the DRT has been widely used for predicting overall intelligibility in voice communication systems and devices. It is the NATO standard and an ANSI standard for evaluating intelligibility of voice coding and communication systems and algorithms.

The DRT tests the discriminability of six distinctive features of consonant phonemes, only. It uses a 2AFC paradigm in which the listener's task with each test token or stimulus word, is to choose between two rhyming words whose initial consonants differ only with respect to one of six features: *voicing, nasality, sustention, sibilation, graveness* and *compactness*. In addition to a total score, the DRT yields more than 24 independent scores. Among these are scores for the discriminability, generally, of each feature, separate scores for each feature state, and various other subscores for each feature, e.g., separate subscores for the discriminability of *sibilation* in voiced and unvoiced phonemes.

3.2 Talkers

The talker sample consisted of three adult males from each of three linguistic backgrounds: American, German and French. They were originally recruited in their native countries by Caldwell P. Smith of the USAF Rome Air Development Center laboratory at Hanscomb AFB, Massachusetts, USA. All, presumably, had formal education in English, but their facility and experience with this language were not independently determined. Each talker recorded several randomizations of the American Diagnostic Rhyme Test words and assorted other speech materials.

3.3 Listeners

Three crews of seven test-naïve listeners, male and female, representing, respectively, American, German and French linguistic backgrounds, were also recruited from present residents of Austin, Texas. None had previous experience with the DRT. All were residing in academic or vocational environments where English was the dominant language of everyday speech communication.

3.4 Testing procedures

The listeners were instructed in DRT testing procedures, given three practice sessions with the test and then presented recorded DRT materials by American, German and French talkers under two conditions, undegraded speech and speech masked by speech-modulated noise at an S/N of 0 dB. The speech materials were presented binaurally over TDH-39 headphones at a comfortable listening level, *circa* 79 dB SPL.

4. RESULTS

DRT results are conventionally expressed in terms of "percent correct, adjusted for chance." In a 2AFC case, the adjustment involves simply doubling the number of observed errors. We will find it convenient to adopt a system of abbreviations for denoting the various permutations of talkers' (TN) and listeners' (LN) linguistic backgrounds: A = English (American), G = German and F = French such that, e.g., GA = German talker(s) * American listener(s), FG = French talker(s)* German listener (s).

Due to the small number of talkers and listeners available for this study, the effects of "talker nationality," "listener nationality" and their interaction are statistically significant in a relatively small number of cases. However, a number of potentially important trends are strongly suggested by these results.

4.1 Results for undegraded speech

Total DRT errors for each of the nine permutations of (TN) and (LN) are shown for the undegraded case in Table 1. Scores were highest when both talkers and listeners were native-born Americans; lowest when the talkers were German and the listeners were French. Listeners of all linguistic backgrounds yielded the highest scores when the talkers were Americans, next highest when the talkers were German and lowest when the talkers were French.

Table 1. Effects of communicators' nationalities on total DRT scores

	Talkers			
Listeners	American	German	French	Mean
American	96.5	92.1	89.9	92.8
German	92.4	89.9	87.3	89.9
French	86.2	82.8	85.3	84.8
Mean	91.7	88.3	87.5	89.2

(For TN, P< .10; for LN, P< .001; for TN*LN, P< .001)

The distribution of voicing discrimination scores for the nine TN * LN permutations are shown in Table 2. Voicing scores were highest when listeners and talkers were American-born; lowest on average when the talkers were of French national origin. Overall, fewest errors occurred with German talkers; most errors occurred with French talkers.

Table 2. Effects of communicators' nationalities on discrimination scores with respect to voicing

	Talkers			
Listeners	American	German	French	Mean
American	97.0	94.9	85.1	92.4
German	90.8	94.3	78.9	88.0
French	89.6	91.7	84.5	88.6
Mean	92.7	93.6	82.8	89.6

(For TN*LN, P< .05)

As shown in Table 3, a consistent positive bias (measured as the difference between "percent correct for the positive feature state" and "percent correct for the negative feature state") appears in all cases involving French talkers, suggesting that French talkers tend to "overvoice". All listeners had a small, but statistically insignificant, tendency to perceive unvoiced phonemes as voiced when the talker was French.

Table 3. Effects of communicators' nationalities ondiscrimination biases for voicing

	Talkers			
Listeners	American	German	French	Mean
American	0.0	0.6	9.5	3.4
German	-0.6	3.0	12.5	5.0
French	-1.8	-4.8	4.8	-0.6
Mean	- 0.8	-0.4	8.9	2.6

Historically, *nasality* has proven to be the most robustly encoded of the six features dealt with by the DRT. Errors were negligible for all talker-listener permutations, but, as shown in Table 4, occurred most frequently with French listeners.

Table 4. Effects of communicators' nationalities on discrimination scores with respect to nasality

-

	Talkers			
Listeners	American	German	French	Mean
American	99.1	99.1	99.4	99.2
German	97.9	99.4	98.8	98.7
French	98.5	96.4	96.1	97.0
Mean	98.5	98.3	98.1	98.3

(For LN, P<.05; for LN*TN, P<.10.)

In all cases, biases with respect to nasality were less than 2%, and no distinguishing trends evident.

Results for the case of *sustention* are shown in Table 5. The main effect for LN is highly significant; the interaction LN*TN is moderately significant. No bias effects approached significance. Here as elsewhere, a significant main effect should be examined critically where an interaction involving that effect is significant. Most of the variation observed here is attributable to cases involving French listeners, the implication of which is that French listeners have greater difficulty than those of other linguistic backgrounds in distinguishing stopped or interrupted consonants from their sustained counterparts. This phenomenon was evident independently of whether the contrasting phonemes involved were voiced (e.g. bat vs. vat) or unvoiced (e.g., pat vs. fat). However, no biases with respect to this feature approached significance.

Table 5. Effects of communicators' nationalities on discrimination scores with respect to sustention

	Talkers			
Listeners	American	German	French	Mean
American	97.6	90.5	94.0	94.0
German	90.2	83.9	89.6	87.9
French	72.0	69.3	78.3	73.2
Mean	86.6	81.2	12.8	14.9

(For LN, P<.001; for LN*TN, P<.10.)

Table 6 shows the distribution of errors with respect to *sibilation.* Errors with respect to this feature were negligible when both talkers and listeners were American, moderate for the case of American talkers and German listeners, but very frequent for all other LN * TN permutations. Moreover, the variation over the nine LN * TN permutations was pronounced, both when the response options involved voiced consonants (e.g., zee vs. thee) or unvoiced consonants (e.g., sing vs. thing). For the voiced case, P<.01 for LN, P<..05 for TN and P<. 05 for the interaction, LN * TN. For the unvoiced case, P<.05 for LN, P<.05 for TN and P<.001 for LN * TN.

Table 6. Effects of communicators' nationalities on discrimination scores with respect to sibilation

	Talkers			
Listeners	American	German	French	Mean
American	98.5	80.4	76.2	85.0
German	93.7	78.0	72.0	81.2
French	81.8	68.8	75.9	75.5
Mean	91.3	75.7	74.7	81.6

(For LN, P<.01; for TN, P<.001; for LN*TN, P<.001.)

Sibilation bias was pronounced in the case of several LN * TN permutations. The extreme negative biases in some cases involving non-American talkers raises the possibility of recording artifacts, but the relatively small biases that occurred in the case of French listeners argues against such an explanation. Bias values for the case of *sibilation* are shown in Table 7.

Table 7. Effects of communicators' nationalities on Bias scores for sibilation

	Talkers			
Listeners	American	German	French	Mean
American	-1.8	- 25.0	-23.8	-16.9
German	-1.8	-21.4	-15.5	-12.9
French	0.6	-7.7	0.6	- 2.2
Mean	-1.0	-18.1	-12.9	-10.7

(For LN, P< .10; for LN*TN, P<.10)

Results for the "place feature," *graveness* is shown in Table 8. Although *graveness* is generally one of the most vulnerable features there is relatively little variability across LN*TN permutations except for that contributed by French listeners, who appear generally to have greatest difficulty in discriminating this feature. This difficulty is evident regardless of whether the critical consonants of the test words were voiced or unvoiced, sustained or interrupted.

Table 8. Effects of communicators' nationalities on discrimination scores with respect to graveness

	Talkers			
Listeners	American	German	French	Mean
American	87.8	90.2	88.7	88.9
German	83.9	85.1	88.1	85.7
French	70.1	77.4	80.7	79.4
Mean	83.9	84.2	85.8	84.7

((For LN, P<.001.)

Table 9 shows the distribution of biases over the nine permutations of LN and TN. Whether due to the characteristics of the talker's or to their own, listeners' responses to the grave test words were biased toward the acute state of the feature in all but two cases, both involving German talkers. This is attributable in part to the fact that four of the items on the grave subtest of the DRT require the listener to distinguish between ʃ and θ, the latter of which is absent from the German phonemic alphabet.

Table 9. Effects of communicators' nationalities on biases with respect to gravèness

	Talkers			
Listeners	American	German	French	Mean
American	-16.1	6.5	-8.3	-6.0
German	-13.1	3.6	-3.6	-4.4
French	-6.5	-17.9	-12.5	-12.3
Mean	-11.9	-2.6	-8.1	- 7.6

(For LN, P < .001)

Table 10 shows the distribution of errors with respect to the place feature, *compactness*. Few errors occurred under with any permutation of LN and TN, only the LN and LN*TN effects approached statistical significance. All biases were negligible in this case.

Table 10. Effects of communicators' nationalities on total scores with respect to the feature compactness

	Talkers			
Listeners	American	German	French	Mean
American	98.8	97.3	95.8	97.3
German	97.9	98.8	96.1	97.6
French	95.2	93.5	96.4	95.0
Mean	97.3	96.5	96.1	95.5

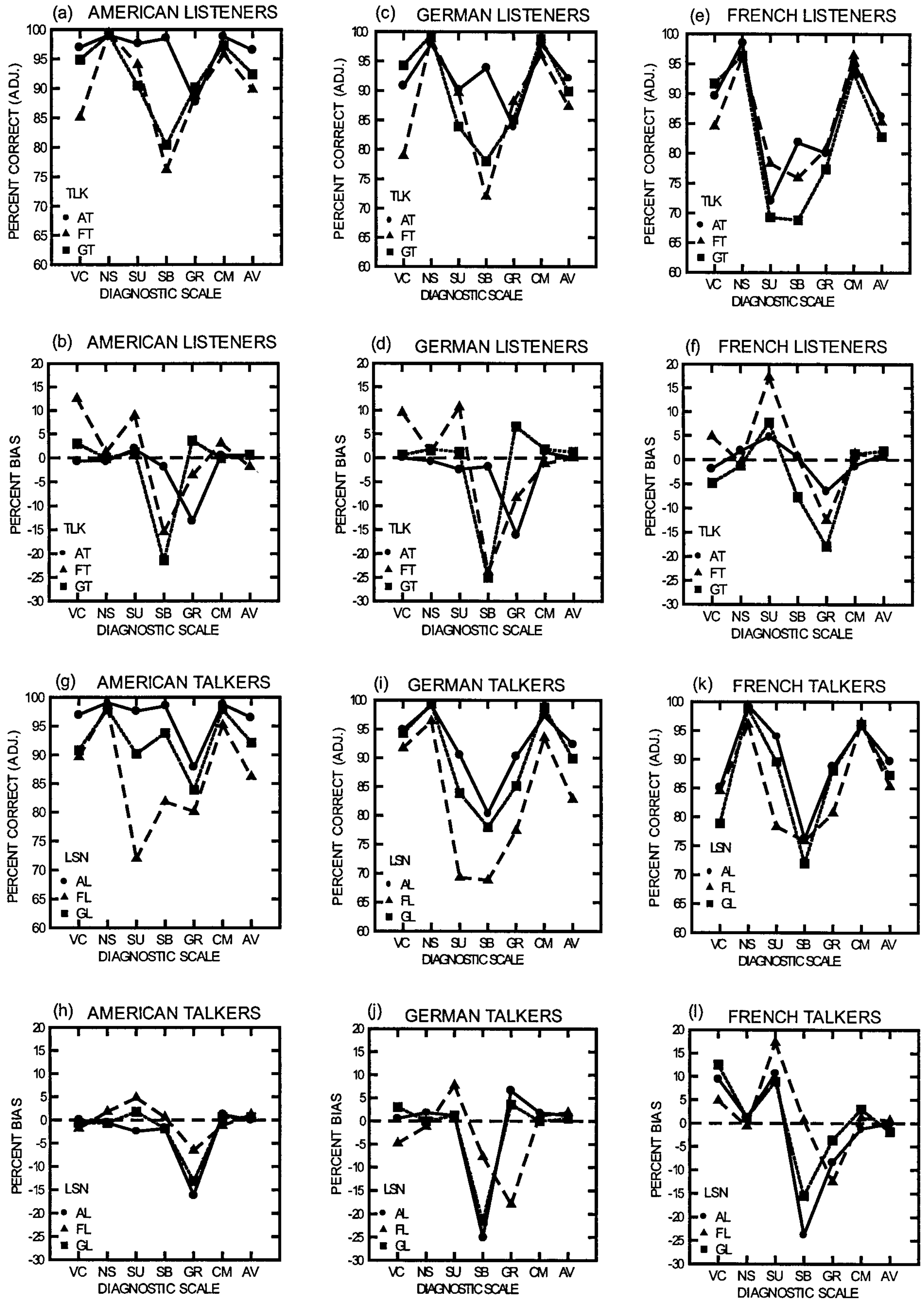


Figure 1. Diagnostic score and bias patterns for the various permutations of listener and talker nationality

4.2. Effects of speech degradation

Recordings of the DRT by the three talker samples were also presented to the three listening crews after being degraded by speech-modulated noise at a speech-to-noise ratio of 0dB. As expected, errors increased significantly across the board. The effects of degradation on total DRT errors are shown in Table 11.

Although significant in two instances, the effects of the communicators' nationalities were generally less pronounced in this case than in the case of undegraded speech, and this trend was generally maintained at the level of individual features. However, when the distribution of errors for the case of degraded speech is compared with that for undegraded speech, differences between the various LN*TN's largely disappear, as shown in Table 12. Evidently, degradation did little to potentiate communication difficulties attributable to specific LN*TN permutations.

Table 11. Effects of communicators' nationalities on total DRT scores under degraded channel conditions (0dB MNRU)

	Talkers			
Listeners	American	German	French	Mean
American	72.3	36.5	37.6	33.9
German	65.0	43.1	41.6	39.9
French	57.9	47.7	45.6	45.1
Mean	65.1	42.4	41.6	39.6

(For LN, P < .01; for TN, P < .05)

Table 12. Increase in error percentages due to speech-signal degradation

	Talkers			
Listeners	American	German	French	Mean
American	24.2	28.6	27.0	26.6
German	27.4	33.1	28.9	29.8
French	28.3	30.5	30.9	29.9
Mean	26.6	30.7	28.9	8

4.3 Relative contributions of listener nationality and talker nationality to communication failures

Figure 1 shows the results of this study from a different point of view. It permits comparisons among patterns of diagnostic scores and biases for the various LN * TN combinations.

Figure 1a shows that, for American listeners, the state of the feature, *voicing*, is most difficult to discriminate in French talkers. In both German and French talkers, *sustention* and *sibilation* are poorly discriminated. Figure 1b suggests that these difficulties are attributable to a tendency of the French talkers to "over voice" and to a tendency of both German and French talkers to "under sibilate."

Figure 1c shows that German listeners had difficulty in discriminating voicing in the case of French talkers and, otherwise, experienced difficulty in discriminating *sustention* and *sibilation* in the speech of their compatriots and that of French talkers. They exhibited a pattern of biases (Fig. 1d) similar to that of American listeners. French listeners had difficulty discriminating the states of all features except *nasality* and *compactness* in talkers of all three nationalities, including their own. They tended to perceive interrupted consonants as their sustained counterparts and to perceive grave phonemes (Fig. 1f) as their acute counterparts.

When the talkers were American, listeners of French origin had serious difficulty discriminating the states of the features *sustention*, and *sibilation*.

When talkers were of German origin, listeners of all nationalities had some difficulty discriminating *sustention, sibilation* and *graveness,* but French listeners had the greatest difficulty in this respect. American and German listeners exhibited pronounced negative biases with respect to *sibilation* but negligible biases in the cases of all other features. French listeners, alone, exhibited a substantial negative bias in the case of the feature, *graveness*.

When the talkers were French, listeners of all nationalities, including French, had substantial difficulty discriminating the states of *voicing* and *sibilation.* Also, French listeners had difficulty discriminating *sustention* and *graveness*. French talkers induced positive biases in *voicing* and *sustention* for listeners of all nationalities; negative biases in the cases of the feature, *sibilation*, for American and German listeners but not for their compatriots.

In the results of ANOVA described above the effects of listener nationality generally proved to be more significant than those of talker nationality. However, an examination of the data from a different point of view provides some potentially important insights. This involved comparing the nine permutations of LS *TN in terms of their error patterns over 224 items of the DRT (including 32 "easy" items. Cluster analysis was the instrument of choice for this purpose. For this case, distance = Pearson r; linkage = complete.

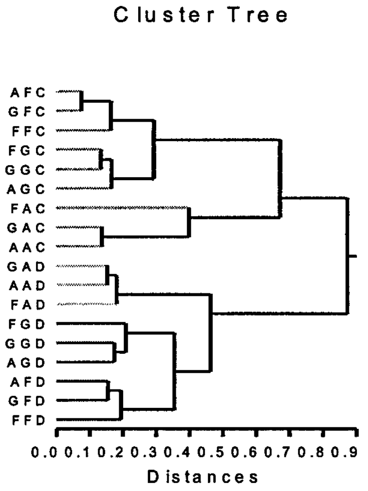


Figure 2. Cluster tree-showing similarities among LN * TN permutations with respect to error patterns across individual DRT items

Figure 2 shows the similarity among the nine permutations of LN and TN in terms of their error patterns under two conditions of signal quality. In the figure, the first letter of the identifying label denotes the nationality of the listeners; the second denotes the nationality of the talkers and the third denotes the quality of the speech signal (C = clear or undegraded; D = degraded).

In the figure, there are two large clusters based on speech signal quality, one containing only the cases of undegraded speech and the other containing only cases of degraded speech. Within each of these, there are three subclusters, all of which are based on the nationality of the talkers. Thus, whereas the nationality of the listener appears to account for the bulk of communication failures, the *patterns* of these failures -- the specific types of error—appear to depend primarily on the linguistic background of the talker.

5. CONCLUSIONS

Subject to the results of additional research, the present findings suggest that remedial programs for non-native speakers of English should place primary emphasis on articulatory rather than perceptual factors in multilingual voice communications. The DRT has potential for purposes of diagnosing communication failures in circumstances requiring communication in English by non-native speakers of English. It may also be a useful tool for evaluating the efficacy of remedial training programs and for evaluating the progress of participants in such programs.

6. ACKNOWLEDGMENTS

The author is indebted to Caldwell P. Smith, whose unpublished work provided the inspiration for this study and who made available recordings of the DRT materials by French and German speakers.

7. REFERENCES

[1] Voiers, W. D. (1977) Diagnostic Evaluation of Speech Intelligibility. In M. E. Hawley (ed.) *Speech Intelligibility and Speaker Recognition*. Benchmark Papers in Acoustics, Dowden, Hutchinson and Ross, Stroudsburg, PA, USA

[2] Voiers, W. D. (1983) Evaluating Processed Speech using the Diagnostic Rhyme Test. *Speech Technology*: 30-39.

SPEECH INTELLIGIBILITY OF NATIVE AND NON-NATIVE SPEECH

Sander J. van Wijngaarden
E-Mail: vanWijngaarden@tm.tno.nl
TNO Human Factors Research Institute.
P.O. Box 23,
3769 ZG Soesterberg,
The Netherlands.

ABSTRACT

The intelligibility of speech is known to be lower if the talker is non-native instead of native for the given language. This study is aimed at quantifying the overall degradation due to acoustic-phonetic limitations of non-native talkers of Dutch, specifically of Dutch-speaking Americans who have lived in the Netherlands 1-3 years. Experiments were performed using phoneme intelligibility and sentence intelligibility tests, using additive noise as a means of degrading the intelligibility of speech utterances for test purposes. The overall difference in sentence intelligibility between native Dutch talkers and American talkers of Dutch, using native Dutch listeners, was found to correspond to a difference in speech-to-noise ratio of approximately 3 dB. The main contribution to the degradation of speech intelligibility by introducing non-native talkers and/or listeners, is by confusion of vowels, especially those that do not occur in American English.

1. INTRODUCTION

Many attributes of individual talkers are known to influence human speech intelligibility. Some of these are at the linguistic level (such as syntactical and lexical aspects [1,2]), some are at the acoustic-phonetic level (e.g. syllabic rhythm and speed, F_0-range, intonation, articulation of different phonemes [3,4,5]). Non-nativeness of a particular talker or listener may be interpreted as a specific category of attributes influencing speech intelligibility.

Among the attributes known to be related to non-nativeness of talkers are vowel-onset time, intonation, speaking rate and phonemic repertoire [e.g. 6,7]. Many fine-grained phonetic studies of second-language talkers have given insight in factors that may contribute to recognition of foreign accents [e.g. 8]. Also, factors contributing to speech intelligibility by non-native *listeners* were investigated [9,10]. Development of accents with experience in using a foreign language has been studied extensively [eg. 11]. Relatively much work has been done in the field of second language (L2) speech perception; however, many studies have been focussed on particular phonetic attributes or phenomena, usually across two (or few) languages.

An important motivation to study the effect of non-native speech, is the effectiveness of human speech communication. From this perspective, it is not important to have detailed knowledge of speech production by L2 talkers; it is more interesting to quantify the effect on the overall speech intelligibility in general terms.

This may be achieved by carrying out speech intelligibility experiments with L1 and L2 subjects (talkers/listeners) in a certain language, in our case Dutch. As with all speech intelligibility tests, a choice has to be made of test fragments: sentences, words or phonemes. In the case of words, meaningful words or nonsense-words may be used. Also, the paradigm will have to be suitable for non-native subjects; on one hand, the limited control of a second language is the object of study, on the other hand it may be experienced as a problem in carrying out some types of speech intelligibility tests (for instance those depending on typing out nonsense words by second-language listeners, who will have a tendency to use native-language spelling of some nonsense words).

2. EXPERIMENTAL SETUP

2.1. Test types

Two types of speech intelligibility experiments were performed: a sentence intelligibility test and a phoneme-intelligibility test based on nonsense-words. The sentence intelligibility test was essentially identical to a standard and widely used test method known as the Speech Reception Threshold (SRT) method [12]. The phoneme intelligibility test is closely related to the equally-balanced CVC test [13].

2.2. Speech Reception Threshold (SRT) method.

The sentence intelligibility test was a standard Speech Reception Threshold (SRT) experiment [12]. This test gives a robust measure for sentence intelligibility in noise, corresponding to the speech-to-noise ratio that gives 50% correct response of short redundant sentences. In the SRT testing procedure, masking noise is added to test sentences in order to obtain the required speech-to-noise ratio. The masking noise spectrum is equal to the long-term spectrum of the test sentences. After presentation of each sentence, a subject responds with the sentence as he or she perceives it, and the experimenter compares the response with the actual sentence. If the response is completely correct, the noise level for the next sentence is increased by 2 dB; after an incorrect response, the noise level is decreased by 2 dB. The first sentence is repeated until it is responded correctly, using 4 dB steps. This is done to quickly converge to the 50% intelligibility threshold. By taking the average speech-to-noise ratio at the ear over the last 10 sentences, the 50% sentence intelligibility threshold (SRT) is obtained.

During the actual experiments, the subjects (listeners) were seated in a sufficiently silent room. A set of Sony MDR-CD770 headphones were used to present the recorded sentences, diotically, to the listeners. Using an artificial head, distortion components introduced by the experimental setup were found to be sufficiently small.

2.3 Semi-open response equally balanced CVC test method

A type of semi-open response CVC (consonant-vowel-consonant) intelligibility test was developed for the purpose of testing phoneme intelligibility with non-native subjects. Using this test, recognition of initial consonants and vowels could be scored, and confusion matrices could be composed [14]. The method is similar to an open-response equally-balanced CVC paradigm [13]. The main differences are that the final consonant is not tested, and that the subject responds by choosing an alternative from a (nearly) exhaustive list of possible CVC-words, instead of typing the word in response to the stimulus. The advantage of this approach is that extensive training of subjects becomes unnecessary, while the construction of confusion matrices is still possible. Problems that were expected using a ‘difficult’ open-response paradigm with non-native subjects were successfully avoided.

During each 3 to 4 minute test, all test phonemes were tested once. Initial consonants and vowels with a frequency of occurrence (based on a Dutch newspaper) below 2% were not included in the test, leaving 17 initial consonants and 15 vowels. Thus, when testing an initial consonant, 17 alternatives were displayed on screen, and for a vowel 15 alternatives. When testing the vowel /ø:/, for instance, the list of CVC words for the listener to choose from could be ‘jaap’, ‘jup’, ‘jeup’, ‘jip’, etc.

In each test, the order of presentation was randomized. The other phonemes in the CVC words, not tested themselves, were selected. Four of these non-tested phonemes, influencing the test through co-articulation effects, were selected per test, in an attempt to maximize the spread of these phonemes over a perceptual space [15]. Several selections of four non-tested phonemes were used for each talker.

2.4. Collection of speech material

The speech material was collected using a B&K type 4192 microphone with a B&K type 2669 microphone pre-amplifier. The sound was digitized using the wave-audio device of a Topline 9000 notebook-computer, which was screened for adequate bandwidth, dynamic range and electronic noise properties This same notebook-computer (with the same audio-device) was used to implement the test procedure.

Since non-native talkers of the Dutch language, matching all criteria, are rather difficult to find, the arguable choice was made to record the material at a location of the talker’s choice. This proved to be an effective measure to facilitate the recruitment of subjects, but lead to a lesser control of the influence of background noise and room acoustics in the recorded material. To limit this influence, the microphone was placed at relatively close range (15 cm). Signal-to-noise ratios were verified to be always higher than 20 dB for all frequencies relevant for speech perception. Hence, no effects of the variation in acoustics and background noise on the outcome of the perceptual experiments is expected.

All speech material was calibrated to have the same speech level for each utterance. In the case of the CVC test, the utterance over which the speech level was determined was not just the CVC-word itself, but also the carrier sentence in which it was embedded.

2.5. Subjects.

Two groups of talkers were recruited, each group consisting of four subjects, two male and two female. The L1 group of talkers consisted of native talkers of the Dutch language without strong regional accents. The L2 group of talkers were native Americans, speaking Dutch fluently but with an accent that was immediately recognized by most listeners.

Perception and production of foreign speech sounds depends on the experience of subjects with the foreign language [11]. Also, the age of acquisition is of importance, leading to a distinction between early and late bilinguals. Generally, the transition age between those categories is found roughly to be puberty [eg. 11,16]. Three of the four L2 talkers had acquired knowledge of the Dutch language above age 23, and spoke Dutch for less than 3 years. The fourth subject (referred to later on as subject L2F8) had first learnt Dutch at age 13 and had been speaking Dutch for 18 years. Although this fourth subject, the only subject that might be categorized as ‘early bilingual’, showed appreciably better control of the Dutch language, the American accent was still readily noticed.

The L1 talkers were selected to match the L2 group in terms of age and level of education.

The L2 *listeners* all had over 12 years experience with the Dutch language (average 20 years), and used the Dutch language frequently in communication at home or work. No special requirements were included in the selection of the L1 listeners.

None of the subjects suffered from speech or hearing impairments, or any unusual hearing loss likely to affect the outcome of test results.

3. RESULTS

3.1. Sentence intelligibility

Four sets of sentence intelligibility experiments were carried out, corresponding to all combinations of L1 and L2 listeners and talkers. The condition with L1 listeners and L1 talkers may be seen as a baseline condition, involving only Dutch subjects. In figure 1, average results are given for these four conditions.

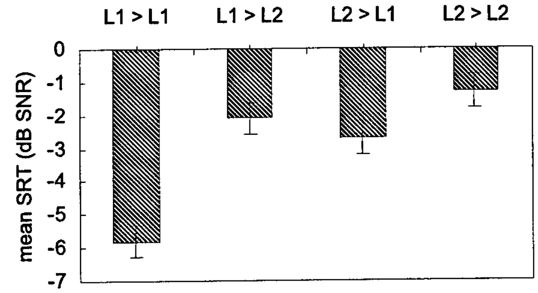


Figure 1. Results for four types of talker-listener combinations (16 talker-listener pairs per condition, mean values and standard errors given). L1 > L2, for instance, means native talker, non-native listener.

The lowest (most negative) SRT value is, as expected, for the baseline group with both L1 listeners and L1 talkers. This means that in this condition the highest noise level may be allowed to still obtain 50% correct sentence responses, down to a speech-to-noise ratio (SNR) of -6 dB.

The condition with L2 talkers and L1 listeners requires a 3 dB lower noise level for the same 50% sentence intelligibility than the L1>L1 condition. The L1>L2 condition (L1 talkers, L2 listeners) also allows less noise for 50% sentence intelligibility; the difference is now nearly 4 dB. The L2>L2 condition, showing the lowest intelligibility results, allows for 4.5 dB less noise.

Figure 1 gives us a general image of the influence of non-nativeness of speakers and listeners on speech intelligibility, at least for these particular L1 and L2 languages. It also shows that, even though the L2 talker group was less experienced than the L2 listener group, having L2 listeners gives relatively more degradation of speech intelligibility than having L2 talkers. The combination of L2 listeners *and* L2 talkers gives an additional degradation which is less than the degradation caused by L2 talkers and L2 listeners separately.

The results of figure 1 are also given in figures 2 and 3, but now by talker instead of talker/listener group. For the L1 listener group (figure 2), all L1 talkers offer better intelligibility than any L2 talker, although the difference between talker L1F4 and L2F8 is not significant. Figure 3 is quite different; to L2 listeners, the highest intelligibility is offered by one of the L2 talkers. The average score by L2 talkers as shown in figure 1 is quite low, but mainly because of talkers L2M5 and L2F6. The difference between L1 and L2 listeners is not as clear with L2 talkers as with L1 talkers.

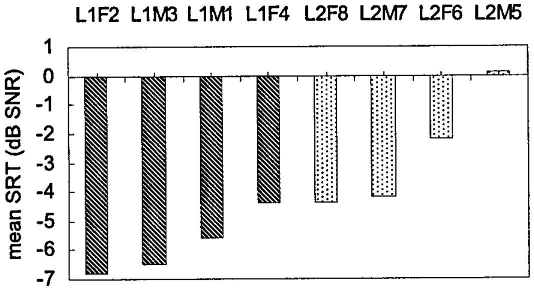


Figure 2. Mean SRT scores for eight individual talkers, with the L1 group of listeners (4 listeners per condition). L2M5, for instance, means L2 talker, male, talker #5.

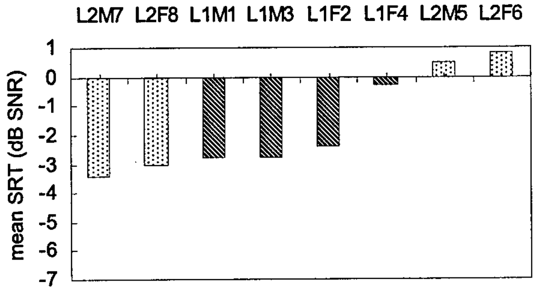


Figure 3. Mean SRT scores for eight individual talkers, with the L2 group of listeners (4 listeners per condition).

3.2. Phoneme intelligibility

The CVC-based phoneme test, although somewhat different in nature, may be expected to yield results that correspond well with the SRT results. However, the CVC test scores are percentages of correctly recognized phonemes, whereas the SRT results are speech-to-noise ratios to obtain 50% sentence intelligibility. To verify correspondence between both test types, CVC experiments were performed at various signal-to-noise ratios. Results, for initial consonants and vowels separately, are given in figures 4 and 5.

Due to the relatively small number of listeners, the experiment data are slightly too noisy for a clear polynomial curve fit. The general trend, however, may well be observed from the data.

At relatively low speech-to-noise ratios, the L2 talker leads to better initial consonant recognition than the L1 talker. At higher speech-to-noise ratios the initial consonant recognition of the L2 talker appears to saturate at a somewhat lower level then the initial consonant recognition of the L1 talker.

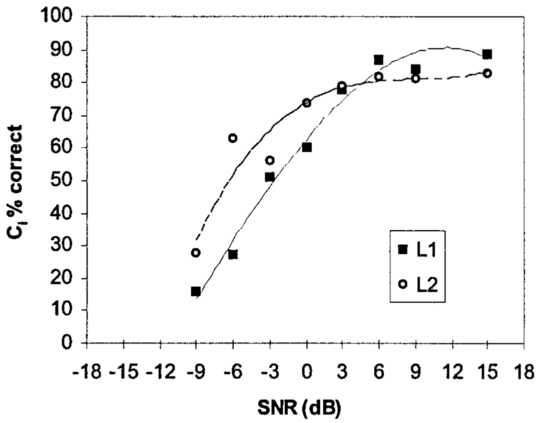


Figure 4. Initial consonant recognition score as a function of speech-to-noise ratio, for a single L1 talker (L1M4) and a single L2 talker (L2M7). Results are mean values for 4 L1 listeners. The lines are third order polynomial fits of the data.

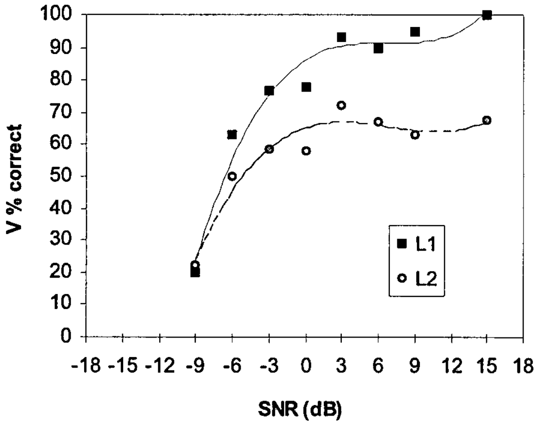


Figure 5. Vowel recognition score as a function of speech-to-noise ratio, for a single L1 talker (L1M4) and a single L2 talker (L2M7). Results are mean values for 4 L1 listeners. The lines are third order polynomial fits of the data.

This is more clearly the case with the vowels; for the L2-talker, vowel recognition saturates at a much lower percentage of correctly recognized vowels. This indicates that, irrespective of speech-to-noise ratio, some vowels by the L2 talker are consistently confused.

At two speech-to-noise ratios (–3 and +15 dB), phoneme recognition was measured for all 8 talkers, with 4 L1 and 4 L2 listeners. Results are shown in figures 6 and 7.

Figures 6 and 7 show, that differences between L1 and L2 speech intelligibility are cause mainly by the vowels. This is in agreement with the data presented in figures 4 and 5.

Non-nativeness of either talkers or listeners has a strong effect on vowel recognition, as may be verified by

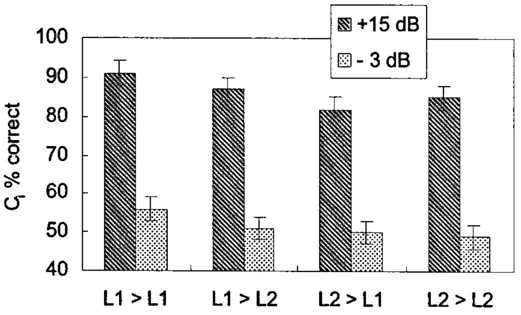


Figure 6. Initial consonant recognition scores at SNR-values of –3 and +15 dB. Results are averages (and standard errors) for 16 talker-listener pairs.

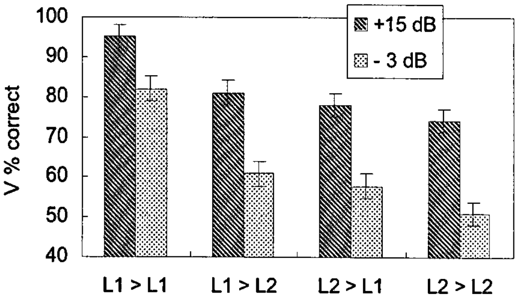


Figure 7. Vowel recognition scores at SNR-values of –3 and +15 dB. Results are averages (and standard errors) for 16 talker-listener pairs.

comparing the L1>L2 and the L2>L1 conditions on one hand, to the L1>L1 condition (baseline) on the other hand. In both cases (L2 talker or L2 listener) the difference in vowel recognition is around 15 percent-points in the +15 dB condition and more than 20 percent-points in the –3 dB condition. This suggests that the effect of additive noise on vowel recognition is somewhat stronger when non-natives are involved.

The loss of vowel intelligibility due to having a L2 talker, is not influenced much by also having a L2 listener. One might hypothesize that a L2 listener would be able to recognize and interpret the L2 accent better, hence recognizing vowels by L2 talkers more effectively. This is not the case, the L2>L1 scores are even slightly higher than the L2>L2 scores. This is consistent with the results from the SRT experiment.

4. ANALYSIS OF VOWEL CONFUSIONS

In order to perform a more diagnostic analysis of vowel confusions, confusion matrices were calculated from the phoneme responses. Although results were obtained at various SNR conditions, only the –3 and +15 dB results included all talkers. In order to obtain sufficiently 'filled' matrices, joint confusion matrices were calculated over both the –3 dB and +15 dB SNR conditions. This way, four matrices were obtained, corresponding to the four L1 and L2 talker-listener combinations. Each matrix contained 32 responses for each vowel (2 SNR conditions, 4 talkers, 4 listeners). Unfortunately, the dataset was insufficiently large to perform meaningful multi-dimensional scaling analyses, which otherwise could have been used to construct 'nativeness-dependent' vowelspaces.

For each of the 15 vowels, in each condition, two types of confusion scores may be calculated from the confusion matrices: the percentage of *false positive* and the percentage of *false negative* responses. A false negative response is the failure to correctly respond with a phoneme upon presentation with that specific phoneme; a false positive response, is responding with that phoneme upon presentation of another phoneme.

The false negative scores are relatively robust, psychophysical indicators of phoneme recognizability; the paradigm is such, that a small false-negative error actually means good phoneme recognition in practice, and vice versa. The meaning of the false-positive error score is different; a large false-positive error may indicate consistent misarticulation of vowels in such a way that they all resemble another vowel; however, it may also reflect a measure of doubt of the listener. Even a vowel that is recognized fairly well as a stimulus, may attract false-positive responses as a response category. Such a response bias may occur, if listeners subjectively classify this vowel as 'difficult' and it as a response to any unrecognized (or similar-sounding) stimulus.

Of the 15 tested vowels, 8 were selected for further analysis. This set of 8 vowels comprised the 5 vowels with the highest overall false-positive scores, and the 5 vowels with the highest overall false-negative scores. The set consists of 6 monophthongs (/ɑ/, /œ/, /y:/, /ɔ/, /o/, /ø:/) and 2 diphthongs (/œy/, /ɑu/). Of this set of vowels, three are not normally found in American English: /y:/, /ø:/ and /œy/. The 8 vowels within the set contribute 64% to the total number of false-negative responses, and 74% to the total number of false-positive responses of all 15 vowels. For the L1>L1 experiment, vowel recognition error scores are given in figure 8.

Note that the false-positive error rate is not limited to a maximum of 100%, since the number of times a vowel is "recognized" when it is not presented is only limited by the total number of vowel presentations.

All error scores in figure 8 are relatively low. The highest percentage of confusions occur with the vowel /o/.

In figures 9, 10 and 11, similar data is given as presented in figure 8, but now for the L2>L2, L1>L2 and L2>L1 experiments.

In figure 9, the distribution of false-negative responses over the vowels is quite different from the distribution of false-positive responses. Remarkably high false-positive scores are observed for the vowels /ø:/ and /œy/, two of the vowels that do not occur in regular American English.

Figure 10 shows a closer correlation between false-positive and false-negative responses than figure 9, with the exception of the vowel /ø:/.

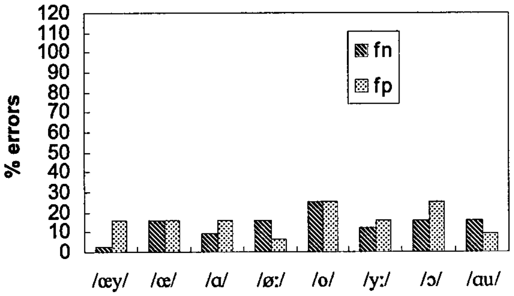


Figure 8. False-positive and false-negative responses in the L1>L1 experiment, to a limited set of vowels. An error score of 100% corresponds to 32 false responses

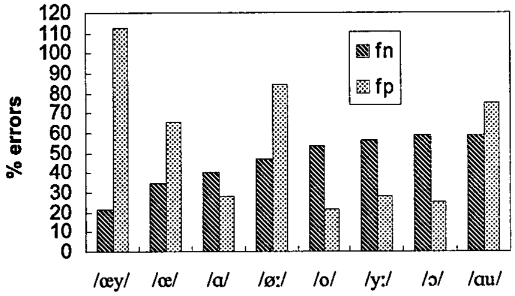


Figure 9. False-positive and false-negative responses in the L2>L2 experiment, to a limited set of vowels.

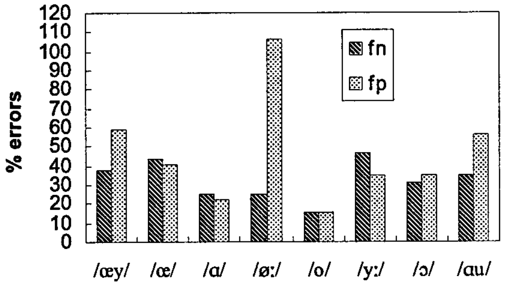


Figure 10. False-positive and false-negative responses in the L1>L2 experiment, to a limited set of vowels.

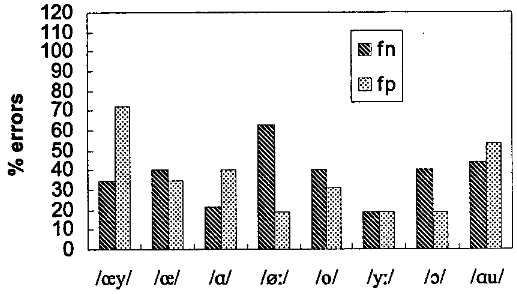


Figure 11. False-positive and false-negative responses in the L2>L1 experiment, to a limited set of vowels.

The vowel recognition errors are considered to be originating largely from two different error sources: non-nativeness of talkers, and non-nativeness of listeners. This is illustrated by the fact that the error scores in figure 8 (only native) are small in comparison to figures 9, 10 and 11.

The highest false-negative score in the L2>L1 experiment is obtained with the vowel /ø:/; this indicates that unusual articulation of this non-English vowel by L2 talkers leads to reduced recognition by L1 listeners. Most of the other vowels also show higher error scores than in the L1>L1 experiment, which indicates that other vowels suffer from unusual articulation as well.

In the L1>L2 experiment, the highest false-negative score is of the non-English vowel /y:/, closely followed by several other vowels. Although the distribution of errors over vowels is somewhat different, the general tendency is similar to the L2>L1 case.

The largest false-positive scores in the L2>L1 experiment are /œy/ and /ɑu/. Many of these responses are given upon presentation of L2-versions of /ø:/, which are usually very close to /œy/ or /ɑu/.

Two vowels, /ø:/ and /œy/, lead to remarkably high false-positive recognition by L2 listeners (L1>L2 and L2>L2 experiments). Not many of the /ø:/ and /œy/ presentations are *missed*, but at the expense of much false recognition. All this reflects the relatively poor model by the L2 listeners of the place of non-English vowels among other vowels.

5. CONCLUSIONS

Two types of speech intelligibility tests (SRT en CVC) produced results that correspond well. Both test types may be used to quantify the effect of non-nativeness on speech intelligibility. The advantage of the CVC test is the diagnostic value of the confusion matrices that may be generated.

Speech intelligibility of L2 (American) talkers of the Dutch language by Dutch listeners is less than L1 (native Dutch) speech intelligibility. The difference corresponds to approximately 3 dB difference in speech-to-noise ratio.

The main cause is consistent confusion of vowels, specifically those that do not occur in American English. This confusion is introduced by L2 talkers, but also by L2 listeners. The total degradation caused by introducing L2 talkers is slightly enhanced (certainly not reduced) by also having L2 listeners.

REFERENCES

[1] Luce, P.A. & Pisoni, D.B. (1998). Recognizing spoken words: the neighbourhood activiation model. *Ear & Hearing*, 19, pp. 1-36.

[2] Hood, J.D. & Poole, J.P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology*, 19, pp 434-455.

[3] Bradlow, A.R., Toretta, G.M. & Pisoni, D.B. (1996). Intelligibility of normal speech I: global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20, pp 255-272.

[4] Sommers, M.S., Nygaard, L.C. & Pisoni, D.B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, 96(3), 1314-1324.

[5] Cox, R., Alexander, G.C. & Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *Journal of the Acoustical Society of America*, 81(5), 1598-1608.

[6] Munro, M. J. (1999). The role of speaking rate in the perception of L2 speech. *Journal of the Acoustical Society of America*, 105(2), p. 1032.

[7] Cutler, A. (1999). Phonemic repertoire effects in lexical activation. *Journal of the Acoustical Society of America*, 105(2), p. 1033.

[8] Magen, H.S. (1998). The perception of foreign accented speech. *Journal of Phonetics*, 26, pp. 381-400.

[9] Bradlow, A.R. & Pisoni, D.B. (1998). Recognition of spoken words by native and non-native listeners: talker-, listener- and item-related factors. Research on spoken language processing , progress report No. 22, pp 74-94: Speech research laboratory, Department of Psychology, Indiana University.

[10] Ingram, J.C.L. & Park, S-G. (1998). Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners. *Journal of the Acoustical Society of America*, 103(2), 1161-1174.

[11] Flege, J.E., Bohn, O-S & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of phonetics*, 25, pp. 437-470.

[12] Plomp, R. & Mimpen, A.M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, 18, pp. 43-52.

[13] Steeneken, H.J.M. (1992). On measuring and predicting speech intelligibility. Doctoral dissertation, University of Amsterdam.

[14] Miller, G.A. & Nicely, P. (1955). An analysis of perceptual confusion among some English consonants. *Journal of the Acoustical Society of America*, 27, 338-352.

[15] Pols, L.C.W. (1977). Spectral Analysis and identification of Dutch vowels in monosyllabic words. Doctoral dissertation, Free University of Amsterdam.

[16] Mayo, L.H., Florentine, W & Buus, S. (1997). Age of second-language acquisition and perception of Speech in noise. *Journal of Speech, Language and Hearing Research*, 40, 686-693.

Report of the plenary discussion on "Human Perception and Assessment"

Chairperson: Edouard Geoffrois (DGA, France)
Reporter: David van Leeuwen (TNO-HFRI, The Netherlands)

Question from *Anderson* for *Voiers*: What is for non-native speech the more important factor in intelligibility: speech production or speech perception? And how is the situation for degraded speech?

Reaction *Van Wijngaarden*: The Dutch /ø:/ is an example in his SRT experiment. The vowel is unfamiliar and difficult for non-natives, but yet there are many false positives with non-native listeners. Because there are also many false negatives with non-native speakers, he argues both production and perception are of comparable importance.

Voiers argues that a non-native effect for listeners is random, while that for speakers is systematic. In order to improve intelligibility, listeners should be trained.

Eklund remarks that linking production with perception is a tricky business. "How much do you listen with your articulators?"

Anderson identifies the question as a chicken-and-egg problem. (Reporter read an answer in the paper: the dioxine chicken was there before the dioxin egg).

Compernolle remembers that he found the difference in perception between the various places of articulation for Indian stops very difficult. Feedback seems necessary in order to learn to differentiate.

Steeneken summarizes the differences in the two perceptual experiments: Voiers measures initial consonant, Van Wijngaarden finds that vowels show a large effect for non-native speech.

Van Wijngaarden adds that also prosody shows an important difference between natives and non-natives, which might be influential on sentence understanding.

Sá Marta reminds us of an experiment conducted by Victor Zue. In noise, vowel recognition is quite robust, while the consonants can only be recognized in broad categories. This is a similar to the experiments of patients using cochlear implants.

Boves replies that Zue's results cannot be extended to continuous speech. He suggests that dynamics of speech are more important than statics in the production/perception of non-natives.

Hunt reports a small improvement seen in Hidden Dynamic Modeling, but Lou believes this is not The Way To Go.

Geoffrois suggests to discuss the question: Is the open software model applicable to speech technology. Can we learn and develop from shared systems, platforms and resources?

Reynolds questions the lifetime maintainability of open software. Resources are not a problem, given LDC and ELRA.

Geoffrois gives the examples of the CSLU toolkit, and the ISIP speech recognition system.

Boves finds open software a tricky issue. He names SPSS as an example. People used to input data until there was a significant effect with no understanding of statistics. Soon, people will think that they can do linguistics. Herman Ney is quoted: "Too many people use off-the-shelf crap software."

Van Compernolle, who was a member in the panel at Eurospeech about the same issue, argues that for a speech recognition system, there are simply too many lines of code in order to make the open software model successful. *Geoffrois* separates users and developers. An example, of the same complexity in terms of lines of code, is the GNU C-compiler. *Van Compernolle* claims that the user base for a compiler is completely different from that of an ASR. The model cannot work.

Jones remarks that the model did not work for Mozilla, the open software version of Netscape.

Reynolds reminds us that in the current, non-open model, licensing is a problem for developers.

Koehler brings up that common tools are important for evaluation etc. For researchers, standardized test databases are also important. Geoffrois replies that the LDC/ELRA model seems to work quite well. *Hunt* notices that the Terrible English Database is quite useless, because no transcription has been made.

Micca says that the exchange of data bases works.

Johnston replies that this is not the case for evaluation, because of learning effects.

Hunt points out that adaptation was ignored in the evaluation. But it works.

Johnston concludes by remarking that in the MIVA project, the exchange of data proved to be extremely successful.

TOWARDS MULTILINGUAL INTEROPERABILITY IN AUTOMATIC SPEECH RECOGNITION

Martine Adda-Decker
Spoken Language Processing Group, LIMSI-CNRS
Bât 508, BP 133, 91403 Orsay, Cedex, France
madda@limsi.fr http://www.limsi.fr/TLP

ABSTRACT

In this communication, we address multilingual interoperability aspects in speech recognition. After giving a tentative definition of multilingual interoperability, we discuss speech recognition components and their language-specific aspects. We give a sample overview of past multilingual speech recognition research and development across different speaking styles (read, prepared and conversational). The problem of adaptation to new languages is addressed. Language-independent and cross-language techniques for acoustic modeling provide a means to port recognition systems to new languages without language specific acoustic data. Pronunciation lexica and text material appear to be the most crucial language-dependent resources for porting. Fast porting being a step towards multilingual interoperability the ongoing efforts of producing multilingual pronunciation lexica and collecting multilingual text corpora should be extended to the largest possible number of written languages.

1. INTRODUCTION

The important progress achieved in speech recognition these last decades has led to successful demos using speech technology. Demos raise expectations when shown to potential users, but yet only few systems are ready for operational use. In a multilingual environment, where potential users have distinct native languages, speech recognition systems have to deal with these different languages or with non-native speaker accents, if a common language is shared. Multilingual environments are common in international communication contexts, which may be political, military, scientific, commercial or tourist contexts. The development of multilingual recognition and spoken dialog systems is hence an important research issue, opening a large spectrum of potential applications. To increase the usability of a prototype system the problems of multilingual and non-native speech have to be addressed efficiently.

Speech recognizers are still very sensitive to non-native speech input or more generally to any kind of condition mismatch. Porting a given system to a new language requires often a significant part of language specific knowledge and resources before achieving viable recognition results. Multilingual corpora have been gathered for language identification and multi-lingual recognition research (OGI-TS, LDC CALLHOME, GLOBALPHONE...). Research and development in multilingual recognition has been widely supported by the European communities (EC) and the Defense Advanced Research Project Agency (DARPA) [39, 5, 12, 40, 14, 43].

In this contribution we address issues of multilinguality and multilingual interoperability in speech recognition.

Using a standard recognizer architecture based an acoustic HMM phone models, pronunciation dictionaries and word N-gram language models, the language-specific aspects of each component are discussed. Many observations are gathered from our experience at LIMSI in developing multilingual speech recognizers [35, 54, 2, 1, 4]. We will then focus on multilingual recognition systems. Without attempting to be exhaustive we try to give an overview of some representative research actions in multilingual and cross-lingual speech recognition.

2. MULTILINGUALITY AND MULTILINGUAL INTEROPERABILITY

There exist about 3000 different spoken languages without accounting for dialects, at the end of this millennium [38]. According to this author only several 100 languages have also a significant written language production for which current speech recognition systems (speech to text systems) are applicable. Studies in automatic speech recognition (ASR) are presently limited to about 20 languages, comprising English, Arabic, Chinese, Japanese, Spanish, French, German, Italian, Portuguese, Greek, Swedish, Danish, Dutch...

Interoperability is a term which is widely used in product marketing descriptions: products achieve interoperability with other products either by adhering to published interface standards (example: the WEB with standards such as TCP/IP, HTTP, HTML) or by making use of a "broker" of services that can convert one product's interface into another product's interface on the fly (example: common object request broker architecture CORBA). Interoperability becomes a quality of increasing importance for information technology products, and naturally, the demand for interoperability of speech technology products arises. Voice over IP (VoIP) protocols have already evolved into world-wide standards (IETF's SIP, ITU,s H.323) to support the emerging voice, data and video services of the next millennium.

For speech recognition systems the term of interoperability is not yet commonly used in the corresponding researcher community. Nonetheless many past or present research actions aim at defining standards for text and speech processing (e.g. the EC EAGLES project on language engineering standards [26]), at developing multilingual resources ([51, 45, 15, 12, 5]), at installing multilingual recognizer evaluations (e.g. the EC SQALE project on multilingual speech recognition evaluation, the DARPA Hub5 program on conversational multilingual speech), and at achieving larger robustness across varying experimental conditions (e.g. the DARPA Hub3 program and Hub4 broadcast news transcriptions). Research towards better multilingual interoperability is supported and fostered by national and international institutions: EC (European Commission), NSF (National Science Foundation), DARPA...

Multilingual interoperability which is the topic of this workshop deals with the problem of designing speech products which are operative in a multilingual context and/or easily portable to new languages. The development of multilingual corpora and resources can be considered as a milestone on the way to multilingual interoperability. Developing such resources however is time-consuming, expensive and their reusability is not always ensured, when moving to new application domains. Important related research areas concern cross-domain portability. Research directions towards more language-independent approaches for speech recognition are also being investigated[47, 32, 31] especially for acoustic modeling.

3. SPEECH RECOGNITION

We briefly review the main components of the recognizer in a statistical approach commonly used for LVSR (*Large Vocabulary Speech Recognition*) [6], [27], [53] and discuss to what extend these components are language-specific. The speech recognizer has to determine the most probable word sequence $\widehat{w_1^N}$ given the acoustic input x_1^T:

$$\widehat{w_1^N} = \arg \max_{\{w_1^n\}} \Pr(w_1^n) \Pr(x_1^T | w_1^n)$$

where w_1^n is a sequence of n words each in the lexicon, n being a positive integer. The acoustic input x_1^T is a feature stream, chosen so as to reduce model complexity while trying to keep the relevant information (i.e. the linguistic information for the speech recognition problem). While the use of language-dependent acoustic features has been investigated (see dedicated session of ICSLP'98) acoustic parameter extraction can be considered as mostly language-independent.

$\Pr(w)$ is to be provided by a language model, and $\Pr(x|w)$ by an acoustic model. The recognition decision is taken as a joint optimization of two terms: $\Pr(w)$, the a priori probability of a word or a word sequence as given by the language model and $\Pr(x|w)$ the conditional probability of the signal corresponding to the word sequence, given by the acoustic model. The output $\widehat{w_1^N}$ is a sequence of items from the vocabulary $\{w_i\}$. Pronounced items which are not in the lexicon (referred to as out-of-vocabulary words or OOVs) are necessarily missing in the recognizer's output, and thus misrecognized. Hence the motivation for maximizing lexical coverage by appropriate definition and selection of the lexical items during training.

- **the acoustic model** $\Pr(x|w)$
  Acoustic units generally correspond to subword units which when compared with word models, reduce the number of parameters, enable cross word modeling and porting to new vocabularies in a monolingual context. For Hidden Markov Model (HMM) based systems acoustic modeling most commonly makes use of context-dependent (CD) phone units.[1] $\Pr(x|w)$ is then obtained via a pronunciation lexicon, where each word w_i is described as a sequence of the appropriate phones:

$$\Phi(w_i) = \phi_1^i \oplus \phi_2^i \oplus \ldots \phi_m^i$$

$$\Pr(x|w_i) \equiv \Pr(x|\Phi(w_i)) = \Pr(x|\phi_1^i \oplus \phi_2^i \oplus \ldots \phi_m^i)$$

  Consistent use of the different phone symbols in the lexicon is probably the most important requirement in pronunciation generation. CD models allow for implicit coarticulation modeling within the acoustic model. Coarticulation due to the surrounding phones necessarily occurs for all languages and hence context modeling should be an effective approach for any language. As CI models merge all different coarticulation effects within the same model, they are more robust as compared to CD models. Separating coarticulation effects using an increasing number of contexts results in a more accurate representation of the acoustic patterns. CD models, accounting for the phonotactic constraints of the language, are hence more language-specific than CI models. Concerning the acoustic phone models (CI or CD) we have to be aware that they always best model the most frequently observed coarticulation effects of the training data. For training corpora with a low lexical variety, CI phone models tend to become word-dependent with possibly poor generalization abilities, both intra and inter language.

  Language-dependent CI models (and even recently context-dependent phone models [31]) have been experimented with for porting a recognizer to new languages.

  To overcome the problem of unobserved sounds when porting acoustic models to a new language, studies aiming at developing multilingual or language-independent acoustic phone models are undertaken both for speech recognition and language identification. Recent researches on language-independent acoustic phone models and cross-language adaptation can be found in [47, 32, 31, 16]. These studies tend to demonstrate the viability of a language-independent acoustic modeling approach. Whereas it is important to be able to bootstrap a recognizer for a new language without prior acoustic models of that language, most researchers tend nonetheless to conclude that using a small amount of language-specific acoustic data either to train language-dependent models or to carry out a language-dependent adaptation, rapidly outperforms foreign language data. MLLR [37] and MAP adaptation techniques are used for adapting cross-lingual or multilingual acoustic models to the new language.

- **the language model** $\Pr(w)$
  Language models are used to model regularities in natural language, and can therefore be used in speech recognition to predict probable word sequences during decoding. The most popular methods, such as statistical n-gram models, attempt to capture the syntactic and semantic constraints

[1] In some real-time systems context-independent (CI) phone units may be used in order to reduce the computation time and search space.

by estimating the frequencies of sequences of n words. The lexical unit, w_i, can be considered the basic observation for statistical language models. The extraction of w_i units from text sources can be more or less straightforward depending on the language (e.g. easy for English or French, difficult in Japanese: no spacing between words)

Given a fixed amount of training data, less reliable language models (LMs) are usually obtained for highly inflected languages (with large lexical variety) than for less inflected languages. The same observation can be made for agglutinative languages. In the latter case decompounding could be applied for lexical unit definition. Tokenizations or text normalizations aimed at reducing lexical variety include some language-independent and a variable amount of more or less complex language-dependent processing [1, 24].

The effectiveness of N-gram LMs for a given language also depends on the validity of the approximation of capturing the language structure within sequences of N words. We know that the validity of this approximation is strongly language-dependent, and hence the N-gram modeling approach will not give the same benefit to speech recognition systems for all languages, even if no limit on available training data were imposed.

- **the decoder** $\arg\max_{\{w_1^n\}}$
  The search space to be explored by the decoder is related to the lexicon size and the language model (LM) complexity. For a bigram LM the search space is proportional to the lexicon size. Pronunciation variants introduce additional entries in the search space. Computational requirements can be controlled by limiting LM size, lexicon size and pronunciation variants.

A speech recognizer should meet the following requirements to guarantee good performance. The vocabulary, the acoustic and language models have to achieve good coverage during the system's operating conditions. The vocabulary should thus contain all or most words likely to appear during operation. This means that the out of vocabulary (OOV) word rate should be minimal. Acoustic models should be able to accurately model the vocabulary words. Context-dependent models allowing for a high coverage of the vocabulary are likely to produce better results, than context-independent models or contextual models which are seldom observed during operation. Similarly language models should produce low perplexity during operation. The same criteria have to be met by multilingual systems.

4. MULTILINGUAL SPEECH RECOGNITION

Ideally a multilingual speech recognizer is able to transcribe speech from different languages, thus identifying both the language used and the word sequence uttered by the speaker. Whereas language and word string can be identified in parallel (multi-lingual recognizer), a more effective way, at least for now, is to prior identify the language using a language identification system on homogeneous acoustic segments, and then decode the word string with the appropriate language-dependent recognizer.

Existing systems have been developed for specific domains and a restricted number of languages, requiring large amounts of annotated language-specific corpora. Without trying to be exhaustive, we can cite some examples of multilingual recognizer developments: the LE-Sqale project on read speech LVSR in English, German and French [35, 54], the DARPA Hub5 program on conversational and multilingual speech LVCSR (*Large Vocabulary Conversational Speech Recognition*) over telephone [9, 12] using SWITCHBOARD and CALLHOME corpora.

4.1. Multilingual LVSR using read speech

The aim of the EC SQALE project (Speech recognizer Quality Assessment for Linguistic Engineering) was to experiment with installing in Europe a multilingual evaluation paradigm for the assessment of large vocabulary, continuous speech recognition systems (LVSR) to assess language-dependent issues in multilingual recognizer evaluation. This project, running from 1993 to 1995 gathered CUED Cambridge (UK), Philips Aachen (Germany), LIMSI Paris (France) and TNO Soesterberg (Netherlands).

In the SQALE project, the same system is being evaluated on comparable tasks in different languages (American English, British English, French and German) to determine cross-lingual differences. The recognizer makes use of phone-based continuous density HMM for acoustic modeling and n-gram statistics estimated on newspaper texts for language modeling. The system has been evaluated on a dictation task developed with read, newspaper-based corpora, the ARPA *Wall Street Journal* corpus of American English, the WSJCAM0 corpus for British English, the BREF-*Le Monde* corpus of French and the PHONDAT-*Frankfurter Rundschau* corpus for German. Experimental results under closely matched conditions are reported. The average word accuracy across all 4 languages is about 85%, obtained for a 20k vocabulary open test (65k open test for German) on a multilingual test set where the OOV rates are kept comparable across languages (about 2% OOVs) Trigram LMs and context-dependent acoustic models were used (about 800 CD models for French and more than 2500 tied-state CD models for English and German). A similar recognizer was developed in Japan [42] using 180M business newspapers. With a 7k vocabulary and an appropriate 7k test set without OOV words, an 80% word accuracy rate is achieved using a bigram LM and about 700 CD models.

In Tab. 1, lexical variety across different languages was investigated for comparable amounts of text corpora[2]. Coverage figures of Japanese reported in [42] are very close to those obtained for Italian. Whereas English achieves the highest lexical coverage (close to 100% for a 65k vocabulary, German has the highest OOV rate of about 5%. For a given speech technology (e.g. a 65k system) better results can thus be expected for English than for German. In German, a major obstacle to high lexical coverage arises from inflected forms and word compounding

[2] The newspaper text corpora compared are the *Wall Street Journal* (American English), *Le Monde* (French), *Frankfurter Rundschau* (German) from the ACL-ECI cdrom, *Il Sole 24 Ore* (Italian), and Nikkei (Japanese).

language	English	Italian	French	German	Japanese
corpus	*WSJ*	*Sole 24*	*Le Monde*	*FR*	*Nikkei*
#words	37.2M	25.7M	37.7M	36M	180M
#distinct	165k	200k	280k	650k	623k
5k cover.	90.6	88.3	85.2	82.9	88.0
20k cover.%	97.5	96.3	94.7	90.0	96.2
65k cover.%	99.6	99.0	98.3	95.1	99.2
20k-OOV%	2.5	3.7	5.3	10.0	3.8
65k-OOV%	0.4	1.0	1.7	4.9	0.8

Table 1: Comparison of *WSJ*, *Il Sole 24 Ore*, *Le Monde* , *Frankfurter Rundschau* and *Nikkei* text corpora in terms of number of distinct words and lexical coverage of the text data for different lexicon sizes. OOV rates are shown for 20k and 65k lexica.

for which morphological decomposition could be effectively applied.

More recently within the German GLOBALPHONE project a multilingual read speech database comprising 15 languages (Arabic, Chinese, Croatian, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil and Turkish) has been collected. Using these data the University of Karlsruhe is working on a multilingual LVSR system [47]. Their research efforts focus on multilingual acoustic modeling and fast bootstrapping of acoustic models for new languages. Speech recognition results have been obtained for 6 languages (word error rates ranging from 10% to near 50%) using 10k vocabularies. Closed test sets have been used by adding missing words in the vocabularies and assigning a low probability to the corresponding monograms in the LM. The multilingual text material is yet too limited for reliable language model estimation.

Experiments in multilingual read speech recognition indicate that good performances can be achieved across languages, provided that sufficient training material is available (10-100 hours of speech, 50-200M of words).

4.2. Multilingual LVCSR using conversational speech

The CALLHOME program [14] (part of the DARPA Hub5 program) was initiated in the US in 1995 in order to study conversational speech between family members over long-distance telephone in a multilingual context. Corpora were recorded in English, Mandarin, Japanese and Spanish (with a variety of dialects) during 1995, Arabic (colloquial Egyptian) and German during 1996. LDC provided the multilingual data to participants. Word error rate results reported in 1997 range from about 40% for English to around 60% for Spanish, Arabic, Mandarin and German. As stated by G. Zavaliakos [55], work on CALLHOME Corpora has verified that current technology is largely language independent. The better results obtained in English can be related to relatively more training data available in this language and maybe a longer and more reliable expertise in English system development. Nonetheless word error rates remain high across the different languages, significantly higher than those reported for read or prepared broadcast speech (around 20% word error rates, Hub4 DARPA program). To measure the impact of mere speaking style on recognition results, by controlling speaker, channel and LM effects, an interesting experiment was carried out at SRI as reported in [14]. Conversational speech was recorded and then transcribed. The same speakers were then invited to read the transcriptions, imitating spontaneous style and a second time in pure read style. Word error rates of about 50% for the true conversational style, drop to about 40% for the false spontaneous elocution, and to around 30% for the read version. Conversational speech doesn't fit the spoken language modeling assumptions as well as read speech (see section 3.). This is particularly true for the articulated phone sequence assumption of the pronunciation lexicon.

Results are consistently disappointing across languages for conversational speech. Whereas read or broadcast speech can be considered as normative to be understood by a large audience, familiar conversational speech spreads a larger variety of individual speaking styles. This may explain the discrepancy observed between performance in read and conversational speech. For the CALLHOME languages about 15 hours of acoustic training data and about 150k words for language model estimation were available. Vocabulary sizes ranged from about 10k to about 20k [12]. Experience taken from conversational speech in English (using Switchboard) shows that significant error reduction (i.e. better conversational speech modeling) can be achieved when moving from 15 to 150 hours of speech and from 150k to 2M words.

4.3. Multilingual Broadcast Transcriptions

The DARPA-Hub4 program, introduced in 1995, concerns broadcast news transcription.

Within the Broadcast transcription program, data collection and corpus design have become more efficient, as large amounts of news are constantly available. Corpus transcription and annotation standards [10] have been developed. Annotated corpora are easily created using freeware transcribing tools [7]. Human broadcast transcription/annotation can range from 10-50 times real-time.

Whereas the main effort is centered on English sources, non English (multilingual) evaluations have been carried out for Spanish [25] and Chinese systems [56], demonstrating the feasibility for other languages. English best results are below 20% word error rate. Error rates on non-native speech (F5 condition [48]) are higher for the corresponding native condition (F0), but the F5 proportion remains low in the overall test sets.

Automatically generated broadcast news transcripts can be used for indexing or document retrieval tasks (NIST SDR program). These research areas go in the direction of speech understanding. The benefits of the Broadcast news task on speech recognition technology progress is discussed in [33].

In Europe the EC is also sponsoring research on multilingual broadcast transcriptions. As an example we can cite the LE4-OLIVE project launched in 1998, which aims to support automated indexing of video material by use of human language technologies and in particular multilingual speech recognition. The prime interest of the OLIVE users is to obtain an efficient, detailed and direct access to their video archives. The users in the OLIVE consortium are two television stations, comprising ARTE (Strasbourg, France) and TROS (Hilversum, Netherlands), as well as the French national audio-video archive, INA/Inatheque in Paris, France, and NOB, a large service provider for broadcasting and TV productions (Hilversum, Netherlands). Technology development and system implementation involve: TNO-TPD (Delft), the project co-ordinator supplying the core indexing and retrieval functionality, VDA BV (Hilversum) building the video capturing software, the University of Twente and the LT Lab of DFKI GmbH Saarbrücken, responsible among others for the natural language technology, LIMSI-CNRS (Orsay, France) and Vecsys SA (Les Ulis, France) developing and integrating the speech recognition modules, respectively.

OLIVE is making use of speech recognition in English, French and German to automatically derive transcriptions of the sound tracks, generating time-coded linguistic elements which serve as the basis for text-based retrieval functionality. Confidence scores are associated with each hypothesized word to allow further processing steps to take into account the reliability of the candidates.

Taking advantage of the corpora available through the LDC, the speech recognizer[18, 21] has been developed and tested on American English. The acoustic models are trained on 150 hours of transcribed audio data, with the language models trained on 200M words broadcast news transcriptions and 400M words of newspaper and newswire texts. Using broadcast data collected in OLIVE, LIMSI has ported its American English system to French. A port to German is underway.

Experiments with 700 hours of unrestricted broadcast news data indicate that word error rates around 20% are obtained for American English. Preliminary experiments in French and German indicate that the word error rates are higher, which can be expected as these languages are more highly inflected than English, and less training data are available. However, it has to be kept in mind, that for the purpose of indexing and retrieval a 100% recognition rate is not necessary, since not every word will have to make it into the index, and not every expression in the index is likely to be queried. Research into the differences between text retrieval and spoken document retrieval indicates that recognition errors do not add new problems for the retrieval task[28].

The broadcast transcription testbed is particularly rich in varying acoustic conditions, topics, domains and languages, with native and non-native speakers. Significant progress in multilingual interoperability can be expected from research in broadcast transcriptions.

4.4. Portability

Porting a speech recognizer to a new language consists mainly in the creation of the language specific acoustic models, pronunciation lexica and language models. As mentioned before the acoustic parameter extraction, the model estimation techniques and the search engine may be considered as language-independent. Porting can thus appear as a rather straightforward process, provided there are sufficient speech and text databases available, together with either a pronunciation lexicon or appropriate letter to sound rules for the pronunciation generation. In the previously described SQALE and CALLHOME programs multilingual resources were provided to the different participants for system development. Porting efforts can then be limited in time to a several months span. Much of the demonstrated progress in speech recognition and spoken language understanding over recent years has been fostered by the availability of large commonly used corpora for system training and evaluation in different languages.

But these resources, while in constant increase are still lacking for many human languages. Especially in military and intelligence applications, interest in exotic languages may arise suddenly and the porting phase should span the shortest duration possible.

4.4.1. Porting using language-dependent resources

In the following we relate some of our experience from the SQALE project where our read speech recognition systems of American English and French have been ported to British English and to German. Language-dependent resources (transcribed speech, text material and pronunciation dictionaries) were available to all partners.

For German the acoustic models were bootstrapped using a mix of French and English models. German acoustic models were then estimated from the PHONDAT read speech database, available for research purposes from the University of Munich. Phondat contains a variety of prompt types including phonetically balanced sentences, a few short stories, isolated letters and train timetable queries. There are a total of 15,000 sentences from 155 speakers. Vocabulary items are rather limited, with only about 1700 different words and the prompt texts are quite different in style from the language model training material (taken from newspaper texts). Despite these relatively mismatched acoustic data as compared to the read newspaper task, and despite the limited amount of distinct lexical items, good recognition performance could be observed for German. But we have to recall two important facts: first the German system used a 65k vocabulary to get acceptable lexical coverage, whereas for the other languages the systems were still using 20k vocabularies. Second the SQALE test sets were designed to achieve similar OOV% rates of about 2% for all languages: the OOV rate with a 20k lexicon without OOV control on the test is 10% in German (2.5% in American English). The OOV problem could be reduced by decompounding compound words, as was done for the numbers during text normalization. Decompounding is however a non-trivial task requiring a refined morphological analysis and even sometimes semantic information. Many compounds can result in two and more items depending on the degree of morphological analysis carried out. For example consider the following compound word occurring in the training texts: *Bundesbahnoberamtsrat* (approximate translation: *Federal-Rail-Head-Office-Chief*). The following decompositions are possible and semantically correct:

Bundesbahnoberamtsrat → *Bundes Bahn Ober Amts Rat*
Bundesbahnoberamtsrat → *Bundesbahn Ober Amtsrat*
Bundesbahnoberamtsrat → *Bundesbahn Oberamtsrat*
Other decompositions such as:
Bundesbahnoberamtsrat → *Bundes Bahnober Amtsrat*
are possible, but semantically poor. This example clearly illustrates that word compounding in German constitutes an OOV-source, as long the recognition system considers a word to be an item occurring between two spaces.

German system development would have taken benefit from a reliable morphological analyzer, both for the quality of the vocabulary (better coverage) and for the LM (more data to estimate Ngrams). As mentioned before even the pronunciations could have been improved, as a lack of consistency may occur when a given morpheme is observed in a long list of compounds.

To conclude here we can say that porting to a new language can be very fast if all resources are available. A baseline system can then be produced in a short delay. In a second step developments can be carried out to better account for language-specificities: typical pronunciation variants, regional accents, stemming, decompounding for agglutinative languages..., Here years can be spent to move away from a baseline performance.

4.4.2. *Lacking training data for the new language: cross-lingual approaches*

A tentative definition of cross-lingual modeling can be the following: resources from one or multiple source languages are used to estimate models for a new target language. Cross-lingual approaches can apply for acoustic phone modeling as similar sounds are often shared across different languages. A relatively large number of research actions aim at defining multilingual or language-independent acoustic model sets [47, 32, 31]. The availability of language-independent acoustic models reduce the problem of lacking acoustic data in the target language.

For lexical and language modeling however language-dependent resources remain mandatory, at least at the present state-of-art. Progress may be achieved through research areas comprising machine translation, multilingual indexing, speech understanding.

The problem of insufficient training material is addressed in [55]. According to this author the dominant factor with respect to performance is the amount of training data available. The author proposes to use the automatically transcribed test data of the new language to adapt the acoustic models to the new language. The proposed method shows a slight but consistent gain in word accuracy when using a subset of automatically transcribed data, selected using a confidence measure criterion, to adapt acoustic and language models.

5. CONCLUSION

We can consider that present recognition systems are potentially multilingual, as the same family of methods and algorithms apply for developing recognizers in a large variety of languages.

Depending on the level of spoken language representation, a more or less important language-dependency is observed. Whereas the acoustic parameter front-end can be considered as mostly language-independent, words and their pronunciations are completely language-dependent. Successful porting to a new target language then requires appropriate language-specific resources, among the most important are text material and pronunciation lexica. The availability and size of these resources is significantly linked to the final recognizer's performance. Developing multilingual resources is expensive, even if dedicated tools exist and speed up the transcription and annotation process. Porting an ASR system to a new target language requires as minimum resource text material for language modeling and pronunciations for the vocabulary. Baseline performance can then be improved either by increasing the volume of training material and/or by adding language-specific knowledge in the various components [52]. Cross-domain research remains an important area, to ensure reusability of these resources when moving to new application domains and to increase ASR interoperability. To overcome the problem of insufficient or missing data researchers are developing interpolation methods to combine corpora. Language specificities, when accounted for properly, will contribute to optimize the recognizer's performance for the new language.

Other research directions concern more language-independent approaches for speech recognition, and more specifically for acoustic modeling. The IPA phone symbol set can theoretically be used to train a collection of language-independent acoustic phone models covering all possible sounds. Language-independent approaches are being investigated [47, 32, 31], and have shown a certain success in porting systems to new languages. Language-independent models have proven useful in bootstrapping recognizers for a new language. Comparative studies show that a small corpus of language-specific acoustic data (1 hour) then rapidly allows to train or adapt better acoustic models [31].

Lexical modeling comprising the definition of the recognizer's vocabulary (word list) with corresponding pronunciations rely on completely language-dependent resources. Vocabularies are often chosen as frequent words occurring in training text corpora which also ensure a good coverage of the application. To overcome a lack of target text corpora for vocabulary definition, bilingual (multilingual) dictionaries can contribute to port vocabularies from source to target languages. But language-dependent resources are necessary for word level modeling (target language text corpora or multilingual dictionaries, letter to sound rules ...). Statistical language modeling for a new target language generally requires huge amounts of text corpora. New challenging research directions joining the domains of machine translation and cross-language information retrieval may contribute in increasing multilingual interoperability in the future.

Multilingual interoperability in automatic speech recognition can be seen as a goal, as a guiding principle to orient research away from purely language-dependent towards more language-independent questions. This is an important goal to strive for. As the number of written languages remains relatively low, we can imagine having baseline resources available for a large proportion of written languages in a near future. An important research issue then consists in defining and developing these resources and generic corpora, which allow for easy adaptation across domains and languages. The availability of these resources for a large proportion of the spoken/written languages will allow to judge the multilingual capabilities of present speech recognition technology. As underlined by V. Zue in his keynote paper of Eurospeech'97 [57], real deployment of spoken language technology cannot take place without adequately addressing this problem of portability.

REFERENCES

[1] G. Adda, M. Adda-Decker, J.L. Gauvain, L.F. Lamel "Text Normalization and Speech Recognition in French", Proceedings of the European Conference on Speech Technology, EuroSpeech, Rhodos, September 1997.

[2] M. Adda-Decker, G. Adda, L. Lamel, J.L. Gauvain, "Developments in Large Vocabulary, Continuous Speech Recognition of German," *IEEE-ICASSP-96*, Atlanta 1996.

[3] M. Adda-Decker, L.F. Lamel, J.-L. Gauvain, G. Adda, *"Activities in Multilingual Speech Recognition at LIMSI'*, CRIM/FORWISS Workshop on Progress and Propects of Speech Research and Technology, Montréal, Oct. 1996.

[4] M. Adda-Decker, G. Adda, J.L. Gauvain, L. Lamel, "Design considerations for LVCSR in French," *IEEE ICASSP-99*, Phoenix mars 1999.

[5] S. Armstrong et al., "Multilingual Corpora for Cooperation", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 975-980, May 1998.

[6] J. Baker, "The Dragon System – An Overview," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol ASSP-23, pp. 24-29, Feb. 1975.

[7] C. Barras et al., "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", Proc. 1st International Conference on Language Resources and Evaluation, Granada, May 1998.

[8] K. Berkling, M. Zissmann, "Improving accent identification through knowledge of English syllable structure", *Proc. ICSLP-98*, pp. 89-92, vol. II, Sidney, Dec. 1998.

[9] J. Billa et al., "Multilingual Speech Recognition: the 1996 Byblos CALLHOME System", *Eurospeech-97*, Rhodos, September 1997.

[10] S. Bird, M. Liberman, "Towards a Formal Framework for Linguistic Annotations", *Proc. ICSLP-98*, pp. 3179-3180, vol. VII, Sidney, December 1998.

[11] B. Byrne et al., "Toward Language-Independent Acoustic Modeling", Summer Research Workshop on Speech and Language, CLSP, John Hopkins University, 1999.

[12] L. Chase, "A review of the American SWITCHBOARD and CALLHOME Speech Recognition Evaluation programs", 1st Int Conf. on Language Resources and Evaluation, Granada, vol II, pp. 789-793, May 1998.

[13] C. Corredor-Ardoy, L. Lamel, M. Adda-Decker, J.L. Gauvain, *"Multilingual Phone Recognition of Spontaneous Telephone Speech," IEEE ICASSP-98*, Seattle, WA, 1998.

[14] C. S. Culhane, "Conversational and Multi-lingual Speech Recognition", *Proc. DARPA Speech Recognition Workshop*, pp. , Arden Conference Center, Harriman, New York 1996.

[15] Ch. Draxler, H. van den Heuvel, H. S. Tropf, "SpeechDat Experiences in Creating Large Multilingual Speech Databases for Teleservices", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 361-370, May 1998.

[16] P. Fung et al., "MAP-based cross-language adaptation augmented by linguistic knowledge: from English to Chinese", *Eurospeech-99*, vol.II, pp.871-874, Budapest, September 1999.

[17] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, "Speaker-Independent Continuous Speech Dictation," *Speech Communication*, **15**(1-2), pp. 21-37, October 1994.

[18] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News: The LIMSI Nov96 Hub4 System," *Proc. ARPA Speech Recognition Workshop*, Feb. 1997, pp. 56-63.

[19] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcribing Broadcast News Shows," *Proc. IEEE ICASSP-97*, Munich 1997.

[20] J.L. Gauvain, G. Adda, L. Lamel, M. Adda-Decker, "Transcription of Broadcast News," *Proc. EuroSpeech'97*, Rhodos, Greece, September 1997.

[21] J.L. Gauvain, L. Lamel, G. Adda, "Partitioning and Transcription of Broadcast News Data," *Proc. ICSLP'98*, Sydney, Nov. 1998, pp. 1335-1338.

[22] P. Geutner et al., "Transcribing Multilingual Broadcast News using Hypothesis Driven Lexical Adaptation", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia 1998.

[23] "Cross-Language Information Retrieval", book by G. Grefenstette, Editor, The Kluwer International Series on Information Retrieval, Kluwer Academic Publishers, Dordrecht / Boston / London, 1998.

[24] B. Habert, G. Adda, M. Adda-Decker, P. Boula de Mareüil, S. Ferrari, O. Ferret, G. Illouz, P. Paroubek, *"The need for tokenization evaluation"*, Proc. 1st International Conference on Language Resources and Evaluation, Granada, May 1998.

[25] J. M. Huerta et al., "The Development of the 1997 CMU Spanish Broadcast News Transcription", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia 1998.

[26] N. Ide, "Corpus Encoding Standards: SGML Guidelines for Encoding Linguistic Corpora", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 463-469, May 1998.

[27] F. Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, **64**(4), pp. 532-556, 1976.

[28] G. Jones, J. Foote, K. Sparck Jones and S. Young, "The video mail retrieval project: experiences in retrieving spoken documents," Mark T. Maybury (ed.) *Intelligent Multimedia Information Retrieval*, AAAI Press, 1997.

[29] F.M.G. de Jong "Twenty-One: a baseline for multilingual multimedia retrieval", *Proceedings of the 14th Twente Workshop on Language Technology (TWLT-14)*, University of Twente, 1998, pp. 189-194.

[30] F. de Jong, J.L. Gauvain, J. den Hartog, K. Netter, "OLIVE: Speech Based Video Retrieval", to appear in CBMI'99, European Workshop on Content-Based Multimedia Indexing, Toulouse, France, October 1999.

[31] S. Khudanpur et al., "Cross-Language Adaptation of Acoustic Models", Summer Research Workshop on Speech and Language, CLSP, John Hopkins University, 1999.

[32] Köhler J., "Language-adaptation of multilingual phone models for vocabulary independent speech recognition tasks", *Proc. IEEE ICASSP-98*, pp. 417-420, vol. I, Seattle May 1998. *Eurospeech'95*, Madrid, Sept. 1995.

[33] F. Kubala, "Broadcast News is Good News", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 83-87, Washington 1999.

[34] L. Lamel, J.-L. Gauvain, "Cross-lingual experiments with phone recognition", *IEEE-ICASSP-93*, vol.2, pp. 507-510, April 1993.

[35] L.F. Lamel, M. Adda-Decker, J.L. Gauvain "Issues in Large Vocabulary, Multilingual Speech Recognition," *Eurospeech-95*, Madrid, September 1995.

[36] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J.J. Gangolf, J.L. Gauvain, *"A Multilingual Corpus for Language Identification,"* Proc. 1st International Conference on Language Resources and Evaluation, Granada, May 1998.

[37] C.J. Leggetter, P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, 9(2), pp. 171-185, 1995.

[38] M. Malherbe, "Les Langages de l'Humanité", Bouquins collection, Robert Laffont editor, 1995 Paris.

[39] J. Mariani, L. Lamel, "An Overview of EU Programs Related to Conversational/Interactive Systems", Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 247-253, Lansdale, February 1998.

[40] J. Mariani, P. Paroubek, Human Language Technologies Evaluation in the European Framework", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pp. 237-242, Herndon, Virginia 1999.

[41] D. Matrouf, M. Adda-Decker, L.F. Lamel, J.L. Gauvain, "Language identification incorporating lexical information," *ICSLP-98*, Sidney, November 1998.

[42] T. Matsuoka, K. Ohtsuki, T. Mori, S. Furui, K. Shirai, "Large Vocabulary Continuous Speech Recognition using a Japanese Business Newspaper (Nikkei)", *Proc. DARPA Speech Recognition Workshop*, pp.137-142, Arden Conference Center, Harriman, New York 1996.

[43] D. Pallett, "The NIST Role in Automatic Speech Recognition Benchmark Tests", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 327-330, 1998.

[44] Rabiner, L.R. and Juang, B.H.: "An introduction to Hidden Markov Models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, **3**(1), pp. 4-16, 1986.

[45] N. Ruimy et al. "The European LE-PAROLE Project: the Italian Syntactic Lexicon", 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 241-248, 1998.

[46] T. Schultz, A. Waibel, "Fast Bootstrapping of LVSR Systems with Multilingual Phoneme Sets", *Eurospeech-97*, pp. 371-374, Rhodos, September 1997.

[47] T. Schultz, A. Waibel, "Language-independent and language adaptive large vocabulary speech recognition", *Proc. ICSLP-98*, pp. 1819-1822, vol. V, Sidney, Dec. 1998.

[48] R. Schwartz, H. Jin, F. Kubala, S. Matsoukas, "Modeling Those F-Conditions – Or Not," *Proc. DARPA Speech Recognition Workshop*, Chantilly, Virginia, pp. 115-118, February 1997.

[49] "Multilingual Text-to-Speech Synthesis - The Bell Labs Approach", book by R. Sproat, Editor, Kluwer Academic Publishers, Dordrecht / Boston / London, 1998.

[50] R. Stern et al., "Specification for the ARPA November 1996 Hub 4 Evaluation," Nov. 1996. *Proc. DARPA Speech Recognition Workshop*, pp. , Arden Conference Center, Harriman, New York 1996.

[51] D. Tufis, N. Ide, Tomaz Erjavec, "Standardized Specifications, Development and Assessment of Large Morpho-Lexical Resources for Six Central and Eastern European Languages" 1st Int Conf. on Language Resources and Evaluation, Granada, vol I, pp. 233-239, May 1998.

[52] U. Uebler, H. Niemann, "Morphological modeling of word classes for language models", *Proc. ICSLP-98*, pp. 1687-1690, vol. V, Sidney, December 1998.

[53] S. Young and G. Bloothooft, Eds., *Corpus-based methods in language and speech processing*, Dordrecht, The Netherlands: Kluwer Academic Publishers, 1997.

[54] S.J. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland, "Multilingual large vocabulary speech recognition: the European SQALE project," in *Computer Speech and Language*, volume 11, nb.1, January 1997,pages 73-99.

[55] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance" Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 301-305, Lansdale, February 1998.

[56] P. Zhan et al., "Dragon Systems' 1997 Mandarin Broadcast News System", *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, 1998.

[57] V. Zue, "Conversational interfaces: advances and challenges", *Eurospeech-97*, vol.I, pp.KN9-17, Rhodos, September 1997 .

MULTILINGUAL VOCABULARIES IN AUTOMATIC SPEECH RECOGNITION

Giorgio Micca[1], Enrico Palme[2], Alessandra Frasca[3]
[1] *CSELT, via G Reiss Romoli, 274, 10148 Torino, ITALY*
[2] *Universita` di Pisa, via Diotisalvi, 2, 56126 Pisa, ITALY*
[3] *Universita` La Sapienza, CSELT, via G Reiss Romoli, 274, 10148 Torino, ITALY*

ABSTRACT

The paper describes a method for dealing with multilingual vocabularies in speech recognition tasks. We present an approach that combines acoustic descriptive precision and capability of generalization to multiple languages. The approach is based on the concept of classes of transitions between phones. The classes are defined by means of objective measures on acoustic similarities among sounds of different languages. This procedure stems from the definition of a general language-independent model. When a new language is to be added, the phonological structure of the language is mapped onto the set of classes belonging to the general model. Successively, if a limited amount of language-specific speech data becomes available for the new language, we identify those sounds which require the definition of additional classes. The experiments have been conducted in Italian, English and Spanish languages. The method can also be considered as a way of implementing cross-lingual porting of recognition models for a rapid prototyping of recognizers in a new target language, specifically in cases whereby the collection of large training databases would be economically infeasible.

1. INTRODUCTION

The design of an Automatic Speech Recognition system for flexible vocabularies requires the definition of an inventory of acoustic-phonetic units reflecting the phonetic and phonotactic structure of a language with the maximum degree of precision compatible with the constraint of statistical trainability of the units. A commonly adopted approach consists in modeling allophones by specifying the phonetic context in which a given phone may appear. The context can extend as far as non adjacent phones ([10]); syllables have been considered as an alternative model, but the larger cardinality of this model – a few thousands of units – prevents from a practical and viable implementation of this approach, even if it shows the benefit of an implicit representation of coarticulation effects. A different method is based on stationary-transitory units [7], where an explicit model is given to transitory segments between two adjacent phones. The higher is the degree of detail in the set of units, the higher is the precision of the model, but the difficulty in the training stage increases correspondingly due to the larger number of units to be trained. For instance, the full coverage of the Italian language in terms of triphones would imply the adoption of an inventory of 7-8 thousands of units, and each of them should appear at least a few tens times in the training corpus to provide enough statistical strength. Transitory-stationary units can be considered as a nice compromise between precision and trainability, because their cardinality is limited to a few hundreds, even if they allow to represent all the most relevant phonotactic phenomena. All these factors are enhanced when we consider the dimension of multilinguality. In this paper we try to give an answer to two questions:

1) how to design a multilingual recognizer for applications requiring the activation of vocabularies including words belonging to several languages; this happens, for instance, in automated vocal access servers providing information on international travel or finance services;
2) how to exploit such a multilingual model in the "interpolation" of a recognizer in a new language, accounting for the similarities of sounds of the target language with respect to the sounds of each language of the multilingual model. The goal here is to base on the robustness and richness of the multilingual model, avoiding the burden of collecting several thousands of utterances from hundreds of native speakers in the target language. This goal clearly impacts the economy of ASR design in applications requiring efficient procedures for cross-language transfer of speech technology.

This research follows two major guidelines:

1) Deployment of cross-language similarity metrics among acoustic-phonetic units, obtaining hierarchies of multilingual sounds;
2) Introduction of the concept of class of transitory unit.

Several different techniques have already been developed for cross-language portability of speech recognition models. In most cases, the starting point is represented by the search for similarities among sounds of different languages. After the pioneering work by Wheatly and al. [1] and the introduction of the concept of poly-phonemes [2], experimented with four European languages (Dutch, British English, German and Italian), several other approaches followed where different combinations of acoustic density clustering and cross-language phonetic lexica mappings were designed and experimented. In [3], we presented an approach similar to [4], and we developed a context-independent multilingual phoneme inventory covering Italian, English, Spanish and German, based on a combination of HMM (Hidden Markov Model) distance measures introduced to compute similarities of acoustic-phonetic units belonging to multiple languages. We also showed how these similarities can be exploited to interpolate acoustic models for a new, undertrained language. In [5] a language-independent approach was attempted, by combining up to eight languages in a global set of polyphones and then by using this model for cross-language transfer purposes. The procedure performs well in the target language, but the large size of the phonetic unit inventory (a few thousands) is an obstacle to a full generalization of the approach towards the direction of language independence. In this paper, we extend the method presented in [3] to transitory units. In [9] a bilingual Italian-German recogniser was investigated, where results from different adaptation schemes are reported.

2. THE METHOD

2.1 Transitory units

Stationary and Transitory units [7] explicitly represent the central, more stable section of phone realizations and the transition from one phone to the adjacent one. For instance, in Italian, the word "bene" ("well"), */b'ε n e/*, is transcribed as

# #b b b'ε 'ε 'εn n ne e e# #

where odd components represent stationary events - # is the silence - and even components represent transitions. *#b* and *e#* are the positional units at the beginning and at the end of the word. In this case, for instance, *b* is the voice bar and *b'ε* represents the transition to the following front vowel. This structure has proven to perform well as far as all the units in the inventory can be properly trained. It may happen that the occurrence frequency of some units is below a minimum threshold in a given language-specific training database; in this case those units would be undertrained. In fact, the minimal statistical coverage requirement can be challenged by the scarceness of data for rare sounds. Furthermore, this drawback is highly emphasized when we look for a global model suitable for multiple languages.

In our method, phones are classified in classes, similar sounds are merged into one class, then these classes are used to build up the set of phone-to-phone transitory classes. Resuming the previous example, the word "bene" can be transcribed as

SL SLVP b VPFV 'ε FVNA n NAFV e FVSL SL

where *SL*, *VP*, *VF*, *FV* and *NA* correspond to "silence", voiced plosive, front vowel and nasal phonetic classes. In [3], the acoustic model was based on context-independent units, therefore all the sounds whose cross-language distance resulted to be below a given threshold were merged into a single class. In the method presented in this paper, classes are introduced only for transitory units. Stationary units are not clustered because they convey the information on the lexical identity of a word. The phonological structure of a given language is preserved, and it is therefore maintained in the multilingual inventory of units. Small classes of transition units preserve a higher degree of acoustic precision than large classes, but reduce the compression factor of multilingual inventories because fewer units are merged into a single transition class. An optimal trade-off between average size of classes and accuracy of acoustic modeling has to be found in order to guarantee a specified level of statistical robustness - trainability - of units without loosing too many details in the model.

2.2 Classes of transitions

Several models were tested according to the design criteria described in the following. We started developing monolingual inventories, then moved to the multilingual case. In the monolingual experiment, we developed two types of unit sets for each of the three languages.

- **Basic class set,** corresponding to the classical taxonomy of consonant and vowel sounds: voiced and unvoiced plosives, nasals, laterals/vibrants, voiced an unvoiced fricatives, affricates; front, central and back vowels. This method produced the inventories *en-170*, *it-114*, *sp-140* for English, Italian and Spanish respectively.
- **Improved class set,** designed according to similarity measures computed on the HMMs of phones in each language. Measures were based on a metrics introduced in [3], where up to five different algorithms are applied to compute the acoustic similarity of the sounds of a language. The phone hierarchy derived from this computation is represented by a dendogram. The data-driven method is as follows: for a given transition of type *xy*, the corresponding transitory class is identified by combining the information provided by the dendogram of both left (*x*) and right (*y*) constituent phones. In this stage, the absolute values of distance measures are taken into account. Two specific classes were introduced for the closure section of plosives (silence or voice bar). The generation of the improved class was carried out in two successive steps. In a first step, the procedure was separately performed for each language and the corresponding HMMs for the transitory and stationary units so obtained were trained. In a first stage, for the English language, we designed the transitory unit classes according to a priori phonetic criteria, and generated the set *en-363-mon*. This model was therefore similar to model *en-170-mon*, but resulted in a finer and more detailed phonetically motivated distinction of classes. Since this model did not yield satisfactory improvements in recognition performance, we moved to the data-driven approach, which produced the set *en-358-mon-dd*. The other two data-driven inventories for Italian and Spanish were it-*220-mon-dd* and *sp-269-mon-dd* respectively. Finally, we obtained the global inventory for the multilingual, multivocabulary model: *mul-670-mul*. It consisted of the combination of the three language-specific sets where classes of different languages, representing cross-language sounds which could be clustered according to the distance measures, were unified. Also closure silences and voice bars were unified across the languages. This cross-language unification operator is represented by the symbol ⊕ in the following formula:

mul-670-mul=en-358-mon-dd⊕ it-220-mon-dd⊕ sp-269-mon-dd

3. EXPERIMENTS

3.1 Speech Databases

Training and test databases used in the experiment consisted of a portion of the SpeechDat databases [8] for Italian and Spanish, while the training English component was collected by CCIR–University of Edimburgh. The size of the databases is given Table 1. Two test data sets were used for English: one from SpeechDat and the other one from CCIR.

	ENG		ITA		SPA	
	Train.	**Test**	**Train.**	**Test**	**Train.**	**Test**
# utt.	**34400**	**1797**	**12800**	**1050**	**5174**	**1730**

Table 1. Training and Test corpora.

3.2 Initialization of transitory HMMs

Two different bootstrapping methods for transitory units were implemented: coarse (c) and fine (f) initialization. With c-initialization, the left state of a transitory unit is given the density function of the rightmost state of the context-independent HMM, represented by a left-to-right, three state topological structure, corresponding to the left component of the transitory unit. The same process is followed for the right state of the transitory unit. Stationary units are assigned the density function of the central state of the corresponding context-independent model. With f-initialization, a Viterbi segmentation of training data is performed using the context-independent three-state models. Acoustic sequences, segmented by the rightmost state of the HMM that correspond to the left component of a transitory unit, are assigned to the left state of this unit; a similar relationship holds for the right state and for the stationary unit. Finally, all segments insisting on a given state are processed by a clustering procedure to derive the Gaussian mixture of the state. The process is iterated on all states of the transitory/stationary unit inventory. Since the segmentation is consistent with the phonotactic constraints (e.g. the leftmost state of transition *xy* is associated to segments of the rightmost state of phone *x* only in contexts where the successive phoneme is *y*), it results in a more precise bootstrap representation of the transitory/stationary units. Anyway, f-initialization requires longer computing time than c-initialization.

3.2 Experimental results

Continuous Density HMMs of acoustic-phonetic units were trained by the K-means algorithm. Each HMM state was represented by a variable mixture density function with up to 32 Gaussians per mixture. The Viterbi decoder generated the N-best scored hypotheses with beam search acceleration. English was the working language for tuning and testing the method; the optimal choices were then extended to Spanish and Italian. Finally the multilingual unit inventory was generated. The multilingual tests were performed on a 535 words vocabulary (475 Italian, 30 English and 30 Spanish). A separate test set for English consisted of a list of 300 railway and underground stations.

3.2.1 Monolingual experiments

The baseline model for English was *en-170-mon*. We tried both c- and f-initialization. Since the latter performed significantly better than the former – Word Recognition rate (WR) of 92.14 compared to 91.59 - we decided to adopt f-initialization in all the successive experiments. The next model, *en-363-mon*, which included quite a larger amount of phonetic knowledge (Section 2.2), brought about only a limited improvement in recognition performance, 7% of Error Reduction rate (ER). Then we moved to the next model, where the new source of information, the distance metrics, was taken into account in the definition of the inventory of phonetic class transitions. Several different mappings of phones to classes were induced by this procedure. Table 2 shows the different allocation in classes for plosive and some fricative/affricate sounds. The new model significantly outperformed the previous one (WR = 94.21, ER = 20.7), indicating that data-driven criteria can be exploited in the optimization of this type of acoustic-phonetic units.

en-363-mon	*en-358-mon-dd*
d t	d t
p b	p k
k g	g
	b
ʃ ʒ dʒ	ʃ tʃ dʒ
tʃ	ʒ

Table 2. Different partitions of English sounds in classes.

The method was applied to Spanish and to Italian languages; this time we directly applied the distance criteria. The error reduction observed with respect to the baseline models was 38.9% for Spanish and 38.8% for Italian (Table 3).

ENG	*en-170-mon*	*en-363-mon*	*en-358-mon-dd*
WR	92.14	92.70	94.21
SPA	*sp-140-mon*	-	*sp-269-mon-dd*
WR	95.55	-	97.28
ITA	*it-114-mon*	-	*it-220-mon-dd*
WR	84.54	-	90.54

Table 3. WR performance of different models.

3.2.2 Multilingual experiments

The multilingual phonetic inventory *mul-670-mul* designed according to the method presented in Section 2 was trained by means of the super-corpus obtained by merging the English, Italian and Spanish training corpora. An example of cross-language alignment of sounds is given in Table 4 for some nasal consonants.

	ENG	ITA	SPA
NA Class	m ŋ n	m n	m n ɲ
GNI Class		ɲ	

Table 4. Classes for nasal sounds.

Interestingly enough, the Italian sound ɲ is not assimilated to the corresponding sound in Spanish, but it is left apart as a single member phonetic class.

These models were tested in two different modes: *monovocabulary* and *multivocabulary*. In the former mode, the test was carried out separately for each language. In the latter mode, all the words of each of the monolingual test vocabularies were merged in a global test vocabulary. The aim of this test was twofold: to probe the preservation of language-specific accuracy of the multilingual models and to evaluate the extent these models might support a multilingual vocabulary recognition task.

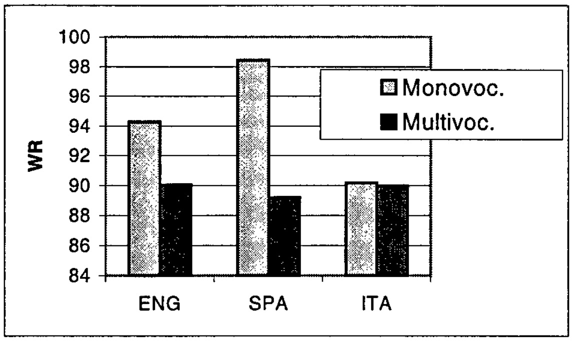


Fig.1. WR performance of *mul-670-mul* multilingual models.

Results are given in Fig. 1.

The test in Spanish privded a significantly better result than the corresponding monolingual model (98.44 WR compared to 97.28). This effect was explained in terms of a greater robustness of multilingual models than the monolingual models, and it was specifically observed in the test with Spanish utterances. In fact the multilingual models took advantage of the larger size of the multilingual training set, and the Spanish language included the smallest language-specific training corpus of the three languages. The tests in the other two languages did not show relevant deviations from the results observed with the corresponding monolingually trained models. The result was WR = 95.23, which is consistent with the figure of 94.68 which was obtained with language-specific models.

A second series of tests was designed and carried out aiming at evaluating the capability of the multilingual models to strengthen the recognition models of a poorly trained recogniser. To this purpose, we selected a portion of about 10% of the training set for Spanish, taking care of including a balanced proportion of male and female speakers. The resulting subset consisted of 517 utterances. WR results are reported in Fig.2. This test clearly points out the effect of strengthening of models for the Spanish language due to the contribution of the training material of the other two languages.

The approach is being evaluated in a cross-language recognition model transfer task involving the Rumanian language. The HMMs of the unseen language will be interpolated by mapping their phonological structure onto the multilingual set of acoustic-phonetic units described in the paper. In this case, since no acoustic data is available in the target language, also the stationary components will have to be bootstrapped from the stationary constituents of the multilingual model. Eventually the HMMs for the new language will be improved by including a limited portion of Rumanian utterances in the multilingual model.

4. CONCLUSIONS

A method for designing multilingual acoustic-phonetic models for automatic speech recognisers has been presented. The approach extends the concept of phone-to-phone transitions in a given language to multiple languages, where similar sounds are represented by a class of transitions. The procedure increases its efficiency and generality as new languages are added to the model. Experiments with a three-lingual recogniser for English, Spanish and Italian languages outline the capability of the model of combining acoustic precision and generalization towards the direction of language independence. The approach is being experimented in a cross-language transfer of acoustic-phonetic knowledge for the Rumanian language.

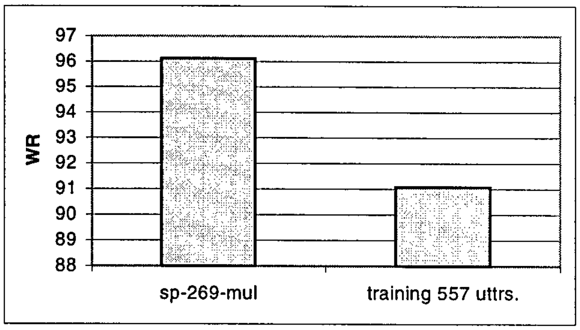


Fig.2. Strengthening of Spanish models.

ACKNOWLEDGMENTS

The authors wish to thank Luciano Fissore for useful suggestions in the experimental stage of this work.

REFERENCES

[1] Wheatley B., Kondo K., Anderson W., Muthusami Y. 1994. An evaluation of cross-language adaptation for rapid HMM development in a new language. Proceedings of ICASSP'94, Adelaide, pp. I-237, I-241.

[2] Andersen O., Dalsgaard P., Barry W. 1993. Data-driven identification of poly- and mono-phones for four European languages. Proceedings of EuroSpeech'93, Berlin, pp. 759-762.

[3] Bonaventura P, Gallocchio F., Micca G. 1997. Multilingual Speech Recognition for Flexible Vocabularies, Proceedings of EuroSpeech'97, Rhodes, pp.355-358.

[4] Köhler J. 1996. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. Proceedings ICSLP'96, Philadelphia, pp. 2195-2198.

[5] Schultz T., Waibel A. 1998. Language Independent and Language Adaptive Large Vocabulary Speech Recognition, Proceedings of ICSLP'98, pp. 1819-1822.

[6] Schultz T., Waibel A. 1998. Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages. International Workshop on Speech and Computer, St. Petersbourg, october 26-29, pp. 207-210.

[7] Fissore L., Ravera F., Laface P. 1995. Acoustic-phonetic Modeling for Flexible Vocabulary Speech Recognition. Proceedings of EuroSpeech'95, Madrid, pp. 799-802.

[8] http://speechdat.phonetik.uni-muenchen.de/, EU Project LE2 4001 SpeechDat.

[9] U. Uebler, M. Schüßler, H. Niemann. 1999. Bilingual and Dialectal Adaptation and Retraining. Proceedings of ICSLP'98, Sydney. Australia, December 1998.

[10] E. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, S. Rieck. 1992. Acoustic Modelling of Subword Units in the ISADORA Speech Recognizer. Proc. of ICASSP'92, San Francisco, CA, Vol. 1, pp. 577-580.

SPEECH RECOGNITION IN 7 LANGUAGES

Ulla Uebler

Bavarian Research Center for Knowledge Based Systems (FORWISS)
Research Group for Knowledge Processing
Am Weichselgarten 7
D-91058 Erlangen, Germany
e-mail: uebler@forwiss.de

ABSTRACT

In this study we present approaches to multilingual speech recognition. We first define different approaches, namely portation, cross-lingual and simultaneous multilingual speech recognition and present results in these approaches. In recent years we have ported our recognizer to other languages than German. Some experiments presented here show the performance of cross-lingual speech recognition of an untrained language with a recognizer trained with other languages. Our results show that some languages like Italian are per se easier to recognize with any of the recognizers than other languages. The substitution of phones for cross-lingual recognition is an important point and we compared results in cross-lingual recognition for different baseline systems and found that the number of shared acoustic units is very important for the performance.

1. INTRODUCTION

Over the years we have studied speech recognition and speech understanding systems in German, and as more and more multilingual applications are needed, the ISADORA system was also used for multilingual speech recognition [1, 8].

The need for multilingual speech recognition applications has risen for example by the growing internationalism like within the European Community or in telecommunications. Thus, applications are developed for recognition in a new language, for example dictation systems are ported to a new language or information systems are developed for e. g. tourist information at airports and train stations which have to be able to understand a couple of languages.

When developing a recognition system for a new language either exclusively for the new language or for the new language in addition to existing languages, the recognition system optimized for the first language has to be adapted to the characteristics of the new language.

During this process, mainly data like the vocabulary, acoustic parameters, language models, and the dialog structure have to be adapted. Most of these adaptations have already been performed before, e. g. when porting a system to a new domain. One topic is still specific to the portation to a new language: the definition and the use of acoustic units. If the recognizer is completely rebuilt for a new language with training material of that language, the definition of new acoustic units arises from the pronunciation of the words in the vocabulary, but when there is not sufficient training material available for the new language or when two languages are recognized at the same time, the acoustic units of the old and the new language have to be set in relation. This problem and solutions to it will be the central aspect in this contribution.

In the following, we will cluster approaches of multilingual speech recognition in order to provide clear definitions for the different approaches and describe characteristics of these approaches. Then we will shortly describe the available data material for our experiments and present different strategies of phone substitution during the transition of languages. We will present experiments and results for different approaches of multilingual speech recognition and phone substitution techniques.

2. DEFINITIONS

When looking at the approaches made in multilingual speech recognition, we find that they may be clustered into three groups depending on the application goal and available data, namely porting, cross-lingual recognition and simultaneous multilingual speech recognition.

When a speech recognition system developed for one language is used for recognition in another language, we speak of *porting.* This step is similar to that of developing an application in a new domain of the same language. The vocabulary and the acoustic units have to be defined for the new language. Special attention must be paid to characteristics of languages like homophones or compound words and other characteristics affecting the recognition process. For these characteristics, algorithms have to be found that can cope with these new problems. The system is then trained with data of the new language. This approach can be found for example in [2, 3, 11].

Another approach follows the same application goal as the approach above with the only difference, that there is not sufficient training material available in the new language. Thus, for *cross-lingual* recognition methods must be found to use training material of another language for a rough modeling of acoustic parameters and only to perform an adaptation with few data of the goal language. One main problem is to determine identical acoustic units or to model existing acoustic units in a way that with few adaptation data a good recognition can be provided. Approaches of this kind can be found for example in [4, 7].

The third cluster of approaches is that of *simultaneous multilingual recognition* . Applications of this approach allow utterances of different languages at the same time for the same recognition system. There are a two main strategies for this approach: firstly, to perform some kind of language identification and perform then monolingual recognition or to have only one recognizer that distinguishes in some way between the languages. For this latter strategy, identical acoustic units may be used across the languages or completely different acoustic units as well as sets of mono- and multi-lingual acoustic units. Also, for language modeling, it may be determined between multi- and monolingual language modeling, which also means that transitions between languages are allowed or not. Approaches for simultaneous speech recognition can be found for example in [1, 8, 10].

3. DATA BASES

The data used in our experiments result from three projects: the EU project SQEL (Spoken Queries in European languages), the EU project SPEEDATA (Speech Recognition for Data-Entry), and from the BMBF project VERBMOBIL.

The SQEL project covers the languages Slovak, Slovenian and Czech in an information system for train and flight time tables. The SPEEDATA project covers the languages Italian and German, both spoken by dialect and non-natives speakers. The task of the project is the entry of land register data in the bilingual region of South Tyrol in the original language, thus the rate of non-native speech will always be around 50 percent. The VERBMOBIL project deals with date scheduling among humans in Japanese, English and German including automatic translation among the languages.

An overview on the training data used from these projects is given in Table 1. With these data, we cover seven languages (German (G1, G2) , Italian (It), Slovak (Sa), Slovenian (Se), Czech (Cz), Japanese(Jp), and English (En)), while German is covered twice. The German data assigned with G1 result from the SPEEDATA project and contain dialect and non-native speakers whereas the data set G2 from the VERBMOBIL project covers only native German speech.

Language	G1	It	Sa	Se
Data/hours	8.6	7.6	5.1	6.1
Distinct vocabulary	5455	6748	1061	955

	Cz	Jp	En	G2
Data/hours	7.2	27.4	9.6	28.5
Distinct vocabulary	1323	3207	2157	7444

Table 1. Acoustic data for each language

The data consist of spontaneous speech for most of the languages, only for G1 and Italian read speech was recorded. Due to the high amount of non-natives and dialect speakers who often try to speak the standard language there are a couple of hesitations and corrections.

The size of the vocabulary differs much among the different tasks and languages. The smallest vocabulary size is observed for the train/flight information domain with around thousand words per language. For the other domains, land register data-entry and date scheduling the vocabulary is higher and varies among 2000 and 7000 words depending on the language. For the experiments we tried to limit the recognition vocabulary to a smaller and equal size for all languages in the experiments without language modeling, but left the original size of the lexicon for the experiments with language models.

4. PHONE SUBSTITUTIONS

Each language has its own characteristic set of phonetic units, and from the phones, different phoneme systems may be built. For example, in Japanese, no distinction is made between /r/ and /l/ and they would thus belong to the same phoneme class in that language, whereas in other languages they are phonemes classes on their own since a semantic difference occurs such that words get a new meaning when e. g. /r/ is replaced by /l/. Some sounds are also unique to some languages, for example the vowel /y/ appears within these languages only in German. If recognition is performed for German with a recognizer that was trained with other languages, the sound /y/ must be modeled although it was not represented in the training material. Thus, the parameters of /y/ must be estimated from other vowels like /I/. Sometimes there is the same symbol used for sounds of different languages, but the acoustic properties differ for these sounds. When recognizing multiple languages simultaneously, it may thus be reasonable to share some sounds across languages and to stay with monolingual units for other sounds.

Thus, for both approaches of cross-lingual and simultaneous multilingual recognition, relations and similarities among sounds of different languages must be found.

In general, we can distinguish between a 1:1 mapping of phones between languages and a n:1 or 1:m mapping of phones, which would mean that for example the parameters of /y/ are estimated as e. g. the mean values of /I/ and /u/. In this work we will refer to the first strategy of a 1:1 mapping. In a rough classification, we distinguish among three different approaches within the 1:1 mapping.

na(t)ive approach: this approach follows the principle a non-native would follow when speaking a second language: he basically has the phonetic inventory of the first language and partially uses that inventory when speaking the second language. Some of the new phones can be learnt by a language learner, but they are not always pronounced correctly, and under stress condition or within difficult words a non-native may fall back to his native phonetic inventory. For example Japanese speaking English or German often confuse the use of /r/ and /l/.

phonetic approach: this strategy follows principles in the production of sounds in the human vocal tract. These characteristics for the production of sounds can be classified into place and manner of production, where the first describes, where obstacles are put in the air flow and which organs are involved in the production of sounds, and the second one describes the manner in which the obstacles act, e. g. a complete or partial closure of the air flow.

Thus, for consonants it can be distinguished with regard to the manner among stop–fricative–approximant–lateral–rhotics and others and for the place between labial–dental–alveolar–palatal-velar–alveolar and others. Another criterion is the voicing of consonants which can be either voiced or unvoiced. For vowels, different tongue positions are distinguished like front–central–back, and for the opening of the mouth among close–close-mid–open-mid–open as well as between rounded and unrounded for the shape of the lips.

The difference between consonants is clearer than between vowels, e. g. a plosive has a complete closure, while others do not have a complete closure, and there is no sound between e. g. a plosive and a fricative. For vowels, the position of the tongue can gradually change and there are transitions between a front and a central vowel, so the distinction and classification of vowels can be more difficult.

For the substitution of sounds in this approach, that sound that agrees in the most phonetic features with the untrained one is taken instead of the unknown one of the goal language. For example, /p/ (plosive, labial, unvoiced) may be replaced by /b/ (plosive, labial, voiced) or by /t/ (plosive, dental, unvoiced). Some hierarchy has to be built in order to define which of the criteria will be changed first.

data-driven approach: this approach determines the similarity among phones with the data given by the trained recognizer. This approach is only possible if there is training data available for the new language, i. e. some adaptation data or for the case of simultaneous multilingual recognition for the decision if acoustic units should be joined.

Measures for the similarity can e. g. be estimated from the Gaussian densities or the codebook parameters of a trained recognizer. Therefore a recognizer must be trained with all languages, and for all observations of a language-dependent sound the similarity parameters like mean values must be estimated and then according to a distance measure the most similar units may be joined. This merging of units can happen in one or more steps and it may also be allowed to split units. The advantage of this approach is that there is no human knowledge or manual work necessary to estimate similarities, but the disadvantage may lie in an exact determination of the segmentation of the speech signal into sounds and consequently an error prone measure for similarities among sounds.

The phonetic description of consonants separates better into classes while measures for the classification of vowels correlate with formant frequencies and of these formant frequencies every compromise between two vowels of, say, 500 and 600 Hertz is possible and thus really different sounds may occur. On the other hand, this characteristic may make it easier to calculate the parameters of sounds by mixing sounds which would average in the same formant frequency.

Another decision is the type of acoustic units that will be used for the target recognizer, especially if the units ought to be mono- or multilingual. For example, to decide for n available languages each containing the sound /a/, if the sound /a/ for the target language (without own training material) shall result from one /a/ of a language or from a mixture of a certain number of /a/'s. With substitution approach 1 and two, the multilingual units may be trained together, and with approach 3 it may be determined according to the data if all or only a couple of /a/'s shall have an influence on the modeling of the new /a/.

Comparing the results of these different strategies for phone substitution it can be found that approaches 1 and 2 are quite similar, of course depending on the priorities set for substitution to manner or place in approach 2. Differences occur mostly when the orthography proposes the pronunciation of another native sound than the similarity according to acoustic features would propose it. For example, in the na(t)ive approach, /u/ may be replaced by /U/ according to the same orthographic spelling [u] rather than to the possibly phonetically closer /o/ if the corresponding criterion is chosen.

Approach 3 is only possible if a certain amount of data is available for all languages; in general it is used for the design of multilingual acoustic units. Errors in this approach can occur if there is not sufficient data available for each language and thus the parameters have not been well estimated. Another source of error for the third approach may be given when the labeling of the speech material according to acoustic units is not completely correct, e. g. with automatic segmentation. Sometimes, silence is assigned to a certain sound and changes this way the statistic properties of this sound.

Another source for errors may be different recording conditions. A consequence may be that sounds of the same language without respect to their phonetic features are estimated as more similar than any sound of the other language. In our experiment, this happened for Slovenian sounds which were for many cases more similar than any sound of another language.

One special phenomenon that has arisen in data-driven decision is the similarity of /j/ and /z/ which have quite different phonetic characteristics (approximant–palatal–voiced vs. fricative–alveolar–voiced) , which has also been shown in several other approaches [5, 6], thus there may be some other measures important besides the phonetic features determined so far.

5. EXPERIMENTS

For our recognition experiments we used the ISADORA recognizer [9] with semi-continuous Hidden Markov Models. We performed experiments both with and without language models, for the experiments without language models we used a reduced recognition vocabulary in order to limit the perplexity of the task.

Instead of the technique of polyphones with context-dependent acoustic units we only used monophones with the phone itself and no context around. The performance decreases by using context-free acoustic units, but only with these units we can hold the number of acoustic units and, even more important, the number of necessary substitutions at a relatively low level.

As baseline systems, we ported our recognition system to the new languages and use the performance obtained with monolingual recognizers for our cross-lingual experiments.

Concerning acoustic units, we considered sounds represented by the same phonetic symbol as identical, and thus, for our cross-lingual experiments, we have to replace those phones whose symbol does not occur in the target language. Furthermore, we did not count replacements for the length of phones, i. e. if there existed only a long vowel like /i:/ and the short correspondent /i/ was needed, we did not count this as substitution. The same is done for Italian geminates, thus /nn/ was set equal to /n/ and the substitution was not counted.

In Table 2 the number of substitutions across languages is shown. There are no substitutions between G1 and Italian since they share proper names of both languages and thus phones of both languages are modeled for each recognizer. Between G1 and G2 there are two substitutions for originally Italian phones (/J/, /L/) which are used in the G1 recognizer. There is a high number of substitutions between the Germanic languages (English, German) on the one side and the Slavic languages (Slovak, Slovenian, Czech) on the other side, once due to the high number of consonants modeled in the Slavic languages and the high amount of vowels in the Germanic languages.

Furthermore, we can observe, that, using the Japanese recognizer for the recognition of any of the other languages, a high number of substitutions has to be made, since the phone inventory of the Japanese language is small in comparison to those of the other languages. On the other hand, for recognition of Japanese with any other recognizer, only a small number of substitutions has to be performed.

Furthermore, we have listed in that table also the number of substitutions for multilingual recognizers, and, of course, the number of substitutions decreases with respect to the corresponding monolingual recognizers, although the complete phone inventory cannot be covered with three languages for all others. We have found out that, besides Japanese, that the phone inventory of the remaining 6 languages can only be covered without substitution only when all 6 languages are involved into training, thus there is no real multilingual inventory possible with a subset of these languages.

Rec \ Lg	G1	It	Sa	Se	Cz	Jp	En	G2
It	0	0	3	1	4	0	8	0
G1	0	0	3	1	4	0	8	0
Sa	10	10	0	4	6	4	12	11
Se	9	9	5	0	7	2	9	8
Cz	12	12	7	5	0	3	11	11
En	11	11	8	3	7	3	0	9
Jp	12	12	9	6	9	0	13	10
G2	2	2	4	0	5	0	7	0
It-G1	0	0	3	1	4	0	8	0
Se-Sa	7	7	0	0	3	2	8	7
Sa-Se-Cz	7	7	0	0	0	2	8	7
G2-En	2	2	3	0	4	0	0	0
G2-En-Jp	2	2	3	0	4	0	0	0

Table 2. Substitution of phones with different languages and recognizers

We performed experiments with na(t)ive and phonetic substitution as well as some preliminary experiments with data-driven substitution for the cross-lingual experiments.

6. RESULTS

The experiments performed for this contribution are done without optimization, i. e. without using the technique of polyphones for acoustic units, without using a polygram verification for language modeling and without optimizing the training procedure in order to obtain recognizers trained at the same level. Thus, the results given here, do not correspond to the optimally trained recognizers, but are comparable to each other with respect to modeling and training.

Results of the experiments with language modeling are given in Table 3 for monolingual and cross-lingual recognition, where the monolingual results are shown in the diagonal. We also give some experiments for multilingually trained recognizers in the second part of that table.

Using different strategies for phone substitution did not lead to significant differences between the na(t)ive and the phonetic approach, but often the na(t)ive approach seems lightly better compared to the replacing strategy proposed by [5]. With data-driven substitution, we found substitutions that correspond roughly to phonetic similarities for Italian and G1 data, but for other languages the similarities do not correspond to phonetic properties. For Slovenian, for example, the phones classified as most similar were in most cases also Slovenian phones, probably the recording conditions dominated over the phonetic similarities.

For all languages besides German G2, recognition is best for the monolingual recognizer trained with data of that language and domain. For G2, recognition showed to be better for the bilingual German-English recognizer under these conditions.

The performance among the languages differs from 37 % for G2 to 94 % for Italian. There are various reasons for this difference: the domains have a different difficulty, in the SPEEDATA task the best recognition is achieved, followed by SQEL and finally the VERBMOBIL task. There are different types of speech and other recording conditions with hesita-

Rec \ Lg	It	G1	Sa	Se	Cz	En	Jp	G2
G1	80.96	87.89	28.05	30.96	55.34	8.49	17.89	20.61
It	94.22	70.74	22.19	38.61	59.03	7.44	18.31	18.70
Slovak	77.60	57.07	88.33	71.03	68.91	7.47	20.15	2.59
Slovenian	86.63	60.66	66.60	90.26	52.25	8.87	30.20	2.01
Czech	81.45	57.07	35.02	58.51	88.57	10.54	22.02	5.85
English	41.41	36.14	35.35	26.77	42.71	48.16	20.28	3.07
Japanese	83.30	56.78	40.70	44.11	36.33	5.48	64.53	1.05
G2	81.53	67.59	39.63	59.40	53.49	12.19	25.19	37.10
G1-It	94.14	86.72	28.19	43.22	63.87	8.81	27.29	19.46
Sa-Se	85.83	60.61	84.23	88.00	62.43	7.41	27.99	1.82
Sa-Se-Cz	86.74	65.02	84.05	85.70	83.88	7.16	29.97	1.53
En-G2	84.40	77.13	38.33	60.71	62.28	24.54	29.60	46.98
En-G2-Jp	87.91	73.89	46.37	64.22	64.42	24.10	52.38	46.50

Table 3. Recognition results for cross-lingual experiments

tions, background noise etc. Furthermore, the size of the vocabulary is different for each language. Finally, the languages themselves differ in the difficulty for recognition, some languages may be easier to be recognized than others due to the phonetic structure, word length and other reasons.

In order to compare the performance of the cross-lingual recognizers trained with one language we averaged the performance of all recognizers besides the one of the original language and domain. Best cross-lingual recognition averaged over the seven other recognizers was achieved for Italian with 78.73 %, worst performance was achieved for G2 with 11.37 %. The ranking in the recognition rate remains the same with respect to the monolingual recognition experiments, only Czech moves one step which could be interpreted that Czech is easier to recognize than Slovenian which moved that step down.

Furthermore, we calculated the ratio of the loss of performance by dividing the cross-lingual performance by the monolingual performance and obtain the same ranking. Here, Italian obtains 85.56 % of the recognition, thus the loss of performance when recognizing with other languages is below 15 % on average, while for G2 with 30.65 % only one third of the performance is achieved.

These both calculations are difficult for interpretation since the similarity of languages and thus the recognizability cannot be taken into account, for example we have two German recognizers in the cross-lingual experiments. Assuming a higher similarity among the Slavic languages, the cross-lingual performance should be higher when recognizing with Slavic recognizers for the Slavic languages than for the others. Furthermore, the cross-lingual recognition of Japanese could be worse because there are no languages similar to Japanese used for recognition.

From these numbers, we can observe, that starting with a poor recognition rate for monolingual recognition, the performance for cross-lingual experiments suffers more than for languages and domains where the performance is already higher itself.

Averaging the performance of cross-lingual recognizers on different spoken languages, we find, that, for monolingually trained recognizers, the best cross-lingual performance was achieved by the Slovenian recognizer which lead three times to the best cross-lingual recognition, whereas Czech, English and Japanese never performed best, thus the Slovenian recognizer seems to be best for cross-lingual recognition in this task. The similarity among languages and therefore their reciprocal cross-lingual performance has a high ranking compared to other languages. Only Slovak and Slovenian showed mutually the best performance for cross-lingual recognizers and may therefore be assumed similar for this speech recognition task, although theoretically, Slovak and Czech should be more similar than those two languages.

For other languages, there is no such symmetry observable, even the two German recognizers do not lead to highest reciprocal results: G1 recognizes best G2, but not vice versa. This may be due to different speaking styles, but more probable to the different speakers, since the speakers of G1 speak with a dialect and with a non-native accent, while the G2 speakers are German natives and do not speak with a strong dialect.

With multilingual recognizers, trained with several languages, performance is worse than with the appropriate monolingual recognizer. Having the target language not included into training, the performance is better than with cross-lingual monolingual recognizers. Unfortunately, for those languages which have the highest cross-lingual performance, no multilingual recognizers were trained, thus often the best monolingual cross-lingual recognizers perform better than the best multilingual recognizers trained in these experiments.

Of the available multilingual recognizers, the G2-English-Japanese recognizer performs best for these data, possibly due to a larger variety in the models provided by Japanese in addition to the Germanic languages models.

7. CONCLUSION

In this contribution, we compared the performance of different monolingual recognizers with respect to cross-lingual recognition. We found with our experiments with non-optimized recognizers (only monophones, no polygram verification in the language models, no optimization in the training), that besides the German G2 task, performance is best for monolingual recognizers. The performance of the different languages differs due to the different difficulty of the task and also due to differing recognizability of the languages.

When monolingual recognition is already bad, cross-lingual performance gets even worse. Thus, for Italian, the average decrease in performance is 15 %, whereas for G2 only one third is recognized with respect to the monolingual recognizer. Cross-lingual performance does not show strong symmetry in the recognition, only Slovak and Slovenian recognize utterances of the other language better than any other language.

When recognizing with multilingual cross-lingual recognizers, performance gets better than with the corresponding monolingual recognizers. Unfortunately, we have not trained all combinations of recognizers, so the combination of the best monolingual cross-lingual recognizers could not always be tested.

Concluding, we found for these languages and domains, that best performance is obtained with monolingual recognizers. For cross-lingual recognition, the choice of the language for training the recognizer is important for the performance. Furthermore, we found that performance increases if training data of more languages are involved and thus both acoustic units are modeled with more variety and more training material as well as more different acoustic units are modeled overall.

REFERENCES

[1] U. Ackermann, F. Brugnara, M. Federico, and H. Niemann. Application of Speech Technology in the Multilingual SpeeData project. In *3rd Crim-Forwiss Workshop*, Montréal, 1996.

[2] J. Barnett, A. Corrada, G. Gao, L. Gillick, Y. Ito, S. Lowe, L. Manganaro, and B. Peskin. Multilingual Speech Recognition at Dragon Systems. In *Proc. Int. Conf. on Spoken Language Processing*, Philadelphia, USA, 1996.

[3] H. Cerf-Danon, S. De Gennaro, M. Feretti, J. Gonzalez, and E. Keppel. TANGORA — a Large Vocabulary Speech Recognition System for Five Languages. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 183–186, Genova, September 1991.

[4] P. Dalsgaard, O. Andersen, and W. Barry. Multi–Lingual label alignment using acoustic–phonetic features derived by neural–network technique. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 197–200, Toronto, Kanada, 1991.

[5] P. Dalsgaard, O. Andersen, and W. Barry. Cross-Language Merged Speech Units And Their Descriptive Phonetic Correlates. In *Proc. Int. Conf. on Spoken Language Processing*, volume 6, pages 2627–2630, Sydney, December 1998.

[6] C.-H. Jo, T. Kawahara, S. Doshita, and M. Dantsuji. Automatic Pronunciation Error Detection And Guidance For Foreign Language Learning. In *Proc. Int. Conf. on Spoken Language Processing*, volume 6, pages 2639–2942, Sydney, December 1998.

[7] J. Köhler. Multi-lingual Phoneme Recognition Exploiting Acoustic-phonetic Similarities of Sounds. In *Proc. ICSLP'96*, Philadelphia, USA, 1996.

[8] E. Nöth, S. Harbeck, H. Niemann, V. Warnke, and I. Ipšić. Language Identification in the Context of Automatic Speech Understanding. In N. Pavesic, H. Niemann, S. Kovacic, and F. Mihelic, editors, *Speech and Image Understanding*, pages 59–68. IEEE Slovenia Section, Ljubljana, Slovenia, 1996.

[9] E. G. Schukat-Talamazzini. *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Künstliche Intelligenz. Vieweg, Braunschweig, 1995.

[10] F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke. A Study of Multilingual Speech Recognition. In *Proc. European Conf. on Speech Communication and Technology*, volume 1, pages 359–362, Greece, September 1997.

[11] S. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.-L. Gauvain, D. Kershaw, L. Lamel, D. Leeuwen, D. Pye, A. Robinson, H. Steeneken, and P. Woodland. Multilingual large vocabulary speech recognition: the European SQUALE project. *Computer Speech & Language*, 11:73–89, 1997.

A MILITARILY OPERATIONAL AUTOMATIC INTERPRETING SYSTEM

*Melvyn Hunt**, Paul Bamberg*, Jay Tucker* & Steven Anderson**

**Dragon Systems UK
Research & Development Ltd
Millbank, Bishops Cleeve,
Cheltenham
Glos, England, GL52 4RW

*Dragon Systems, Inc
320 Nevada Street
Newton, MA 02160
USA

ABSTRACT

This paper describes a real-time interpreting system in which the operator speaks one of around 4000 phrases in one language, which is automatically recognised and the corresponding spoken phrase in the target language is played through a loudspeaker. This system has been used operationally by NATO forces. The basic system is first described, followed by an account of the wide range of uses to which this relatively simple one-way interpreting system can be put. Some developments of the basic system are then listed, both developments that are already in place and some that have potential for future implementation. Finally, an account is given of some relevant research on the use of statistical phonetic mapping techniques for extending the usability of such systems to non-native speakers of the source language.

1. BACKGROUND

This paper is concerned with a cross-language automatic interpreting system that has been used operationally by Nato forces. Experience with this system demonstrates, we feel, that relatively simple technology can perform a surprisingly useful task.

Reportedly, during the Gulf War Alliance forces had some difficulty in finding enough Arabic speakers to communicate with the very large numbers of Iraqi prisoners that had to be handled. This led the US military to seek some automated means of communicating with people in languages other than English. Dr. Lee Morin of the U.S. Navy created the "Medical Translator," with a point-and-click interface, to permit simple medical interviews. Anticipating a similar situation in Bosnia, DARPA asked Dragon System to add a voice interface to the Medical Translator, and Dragon Systems responded by developing the Multilingual Interview System, which first saw operational use in Serbo-Croatian with US forces assigned to the UN in Bosnia. More recently, in response to the troubles in Kosovo, an Albanian version has been developed.

2. THE MULTILINGUAL INTERVIEW SYSTEM

The first version of this system, which saw service in Bosnia, was based on the discrete-utterance, large-vocabulary speech recognition system, *DragonDictate*® [1, 2]. The term "discrete-utterance recogniser" is normally taken to be a synonym of "isolated-word recogniser". However, DragonDictate is capable of accepting and recognising long phrases. The first system developed had a "vocabulary" of 4000 such fixed phrases. They could be developed simply by providing the orthographic text of for each phrase and using the built-in 200,000-word pronouncing dictionary to develop a phonetic spelling for the phrase. Each phrase, no matter how long, was modeled as if it were an "isolated word."

The corresponding phrases in the target language are recorded by a native speaker of that language and stored as digitised waveforms. This provides a spoken output that is much more intelligible and natural than is possible with the current state-of-the-art in automatic text-to-speech systems. In any case, text-to-speech systems, or at least good-quality text-to-speech systems, exist only for a small number of major languages, not necessarily including the languages of interest for the Multilingual Interpreting System.

An operator speaks one of the 4000 phrases. He or she then confirms that the recogniser has correctly identified the phrase, either by seeing it displayed on a screen or — if eyes-free operation is needed — by having a recorded version of that phrase spoken back to the user. After confirmation, the phrase is then converted to the target language by simple table look-up, and the corresponding recorded phrase in the target language is played out through a loudspeaker. In cases where there exist several phrases that differ only in their final words, the system saves disk space by playing back a concatenation of two or more recordings, *e.g.* "I am a member" + "of the NATO peacekeeping forces."

Because the discrete-utterance recogniser needs to perform less computation than a continuous speech recogniser, the hardware requirements are more modest. Portable computers using 486-style processors can be used in place of the Pentium-style processors needed for large-vocabulary continuous speech recognition, reducing the weight and our requirements of the portable equipment. The only technical weakness of the system is that it is vulnerable to errors in the "rapid match" portion of the discrete-utterance recogniser, which narrows the list of candidate phrases to 1000 or fewer by inspecting only the first 300 milliseconds of speech.

One of the operational systems was based on the *Fujitsu* portable PC, which thanks to speech recognition was particularly compact in that it needed no keyboard or mouse during use, its only input being via a headset-mounted microphone. This PC has the unusual feature of a monochrome transflective display that is easily readable in bright sunlight.

Although the system could be used with all 4000 phrases simultaneously active, it has often been found to be convenient to use the phrases in situation-specific subsets, such as those appropriate for a medical examination or for landmine clearance. Even the most dedicated user could not memorize all of the 4000 phrases, but individual users quickly learned the subset needed for their own tasks. The system included several techniques (categories, keyword search, prebuilt dialogues) to help users find phrases that were unfamiliar to them.

When used as a conventional dictation system, *DragonDictate* normally needs to be adapted to the voice of the user. However, in the Multilingual Interpreting System, especially when used with phrase subsets, it has generally been found to perform satisfactorily in speaker-independent mode, allowing military personnel of the same gender to share the same system freely without any need to signal to the system that a change of user has occurred.

Although this system was originally developed for operators whose language is American English, it could in principle be operated in several other major languages, since the *DragonDictate* recogniser on which it is based is available in British English, French, Italian, German, Spanish and Swedish. In practice, because of the length of the phrases and the vocabulary restriction, speakers of British English and indeed other national variants of English can satisfactorily operate a system set up for American English.

Generating a system for new target language is a simple operation, requiring the 4000 phrases to be translated into the new language and a native speaker of that language to record them.

3. USES OF THE MULTILINGUAL INTERVIEW SYSTEM

The system described in the previous section clearly operates only in one direction: from English into Serbo-Croatian, for example. This might appear at first to be a crippling limitation, since spoken communication is normally a two-way process. There are, of course, situations, such as crowd control, where one-way communication is all that is needed. But in a surprisingly large proportion of cases where two-way communication is needed, the Multilingual Interpreting System can perform a useful task. This section will describe just a few of the ways and situations in which it can be effective.

Often, questions can be posed in a way that allow yes/no responses. The military user can learn the words for "yes" and "no" in the target language or head movement gestures for these words may be common between the two languages. It is of course important to avoid negative questions (*e.g.* "You aren't injured, are you?) where the meaning of responses using the words normally translated as "yes" and "no" differs between languages.

In medical examinations, many things that need to be said are either instructions (*e.g.* "Please lie still", "Please open your mouth") or items of information (*e.g.* "I'm going to give you an injection to ease the pain"). The appropriate response to some others (*e.g.* "Point to where it hurts") is a gesture rather than a verbal response.

In gathering personal information, the individual being addressed can respond by writing down an answer (*e.g.* his or her date of birth, name, place of birth...). This will always be comprehensible for numerical information and for other information provided the language uses the Roman alphabet. Even when it does not use the Roman alphabet, the written information can be saved as a bitmap and taken away to be interpreted by others not necessarily located in the field of operation.

In the important area of avoiding or clearing minefields, individuals providing information can indicate locations on a map displayed on the computer screen and point to the kind of mine that has been laid when shown a screen that displays pictures of various mines. Such responses can be acted upon immediately or saved as annoted graphics for later review.

At security checks, the phrases being interpreted will normally be instructions, such as a request to leave a vehicle, to present identity papers or to hand over any firearms.

Finally, in cases where what is required is an extended verbal response, but the information required is not urgent, the person being interrogated can have his or her spoken response recorded for translation away from the field of operation. The Multilingual Interview System is provided with the ability to make such recordings.

One of the advantages of the Multilingual Interview System surprisingly did not involve communication in the conventional sense at all. Reportedly, the use by soldiers of the system was a source of fascination to Bosnian young men, who were drawn into better relations with the soldiers because of it.

4. DEVELOPMENTS

The second version of the Multilingual Interview System allowed the recognition process to be enhanced from the fixed-phrase recogniser used in the first version to a true continuous large-vocabulary recogniser, namely the recogniser used in Dragon's general-purpose continuous speech recognition product, *Dragon NaturallySpeaking*™ [3]. At the price of requiring a more powerful microprocessor, this allows much greater flexibility in the form of the input in the source language. For example, a user does not have to remember whether "Please point to where it hurts" or "Point to where it hurts, please" is the required form of the phrase: both can be accepted with the more flexible arrangement that the continuous recogniser permits.

Both versions of the Multilingual Interview System have thus been based on a recogniser designed primarily for the very large vocabulary speech recognition needed for general-purpose dictation, where the grammar used must be probabilistic and allow in principle any word to occur in any context. Dragon Systems have recently developed a more compact recogniser suited to tasks in which the vocabulary and the structures of the phrases constructed from the vocabulary are more constrained. This will in principle allow the Multilingual Interview System to function on simpler hardware with lower power consumption yet with the flexibility provided by the second version just described.

Future developments can be envisaged that take advantage of research carried out at Dragon Systems on robust speech recognition in noisy conditions [4]. Currently, the operator of the Multilingual Interpreting System wears a headset-mounted close-talking microphone. Such a microphone contains a pressure-gradient element, making it relatively insensitive to distant sources of noise and consequently able to function well in high-noise environments. In some situations, however, it may be more natural and convenient to use a more conventional desk- or lapel-mounted microphone. Such microphones do not have the noise-cancelling properties of the headset-mounted microphone, but the developments in noise-robust recognition should allow the Multilingual Interpreting System to function with them in noisy conditions, at least when the noise is reasonably steady, such as noise from machinery.

The work on speech recognition in noise has also led to techniques for very rapid adaptation to the voices of native speakers of the source language and to techniques for compensating for the Lombard effect [5, 6] (the changes that occur in the voice of a user when the noise environment changes — principally an increase in the loudness of the speech when the noise gets louder).

5. POSSIBILITY OF USE WITH NON-NATIVE SPEAKERS

The performance of a system with non-native speakers is clearly of central interest to this workshop. Although no research using the Multilingual Interview System has been carried out with non-native speakers, relevant tests have been carried out in the framework of the development of speech recognition in noise just described [4].

The performance of the noise-robust recognition system was tested in noisy conditions in speaker-independent mode with both native and non-native speakers in a phrase recognition task not dissimilar from that in which the Multilingual Interview System might be used with a vocabulary of a few hundred words. The non-native speakers showed error rates roughly six times greater than the native speakers. The phonetic models used in recognition were then adapted using data from prompted training utterances from the speakers. There is no attempt to train the recogniser for specific words; rather, we use statistical techniques to map [7] the speaker's phonetic system into that of the standard language. In these particular tests, the training utterances were not chosen to give a balanced phonetic coverage of the task vocabulary but rather were selected randomly from phrases that can occur during use.

Figure 1 shows that adaptation is very effective with the non-native speakers, with a useful reduction in error rate after just 10 utterances, and a factor-of-three reduction after 80 utterances. The proportional reduction with native speakers is much less: only about 40% for female speakers and no clear improvement at all for male speakers. We have found in our tests that adaptation with our non-native test speakers always improved recognition performance, while with native speakers performance could even be degraded if an inadequate amount of adaptation material was used. This experience encourages our belief that the statistical techniques are indeed mapping acoustic realisations of phonemes from some consistent but non-standard forms produced by non-native speakers to something closer to standard forms. Note, however, that despite the evident effectiveness of the adaptation, the error rate with the non-native speakers remains about twice as high as the unadapted error rate with native speakers.

Of course, the term "non-native speaker" covers an immense range of deviation from the standard form of the language, both in extent and in the type of deviation. Nevertheless, the adaptation behaviour seen here might reasonably be expected to be seen with any non-native speakers whose deviations from the norm correspond substantially to non-standard but consistent acoustic realisations of particular phonemes.

A key aspect of this kind of adaptation is that it is "supervised"; that is, the system knows what the speaker actually said. In the experiments just described this was achieved by prompting the speakers to produce the training utterances. In the Multilingual Interpreting System it is usual for the user to confirm that the system has correctly recognised the phrase spoken before it is translated into the target language. This process achieves the end of confirming to the system what the speaker actually said, and consequently adaptation of the form just described could be carried out unobtrusively during use.

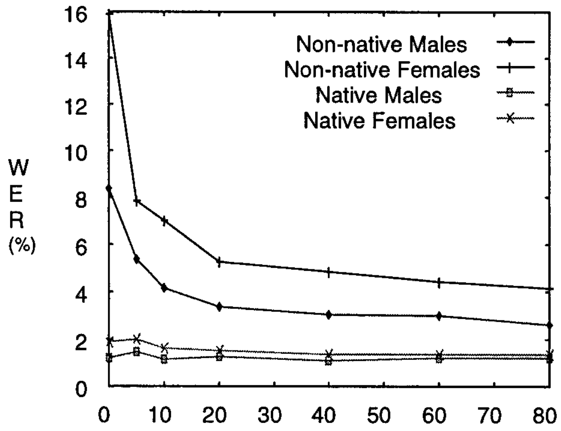


Figure 1 Word recognition error rate versus number of utterances used to adapt to the speakers for native and non-native speakers

6. CONCLUSIONS

This paper has attempted to show that an operationally useful automatic interpreting system for both military and non-military applications can be constructed from the current widely available large vocabulary speech recognition technology. The one-way nature of the interpreter does not prevent it from being effective in a wide range of tasks. There is much scope for the development of such a system to allow use with non-native speakers and in noisy environments without close-talking microphones, and for further reductions in the size and weight of the hardware required.

REFERENCES

1. J. Barnett, P. Bamberg, M. Held, J. Huerta, L. Manganaro, A. Weiss, "Comparative Recognition Performance in Large-Vocabulary Isolated-Word Recognition in Five European Languages", *Eurospeech '95*, vol. I, pp. 189-192.

2. J. Barnett, A. Corrada, G. Gao, L. Gillick, Y. Ito, S. Lowe, L. Manganaro, B. Peskin, "Multilingual Speech Recognition at Dragon Systems", *ICSLP '96*, pp. 2191-2194.

3. R. Roth, L. Gillick, J. Orloff, F. Scattone, G. Gao, S. Wegmann, J. Baker, "Dragon Systems' 1994 Large Vocabulary Continuous Speech Recognizer", *Proc. ARPA Spoken Language Systems Tech. Workshop*, Austin, 1995, pp. 116-119.

4. M. J. Hunt, "Some Experience in In-Car Speech Recognition" *Proc. IEEE/Nokia Workshop on Robust Methods for Speech Recognition in Adverse Conditions.* May 25-26, 1999, Tampere, Finland, pp. 25-32.

5. E. Lombard "Le Signe de l'Elevation de la Voix", *Ann. Maladies Orielle, Larynx, Nez, Pharynx*, Vol 37, 1911, pp. 101-119

6. J-C Junqua, "The influence of Acoustics on Speech Production: A Noise-Induced Stress Phenomenon Known as the Lombard Reflex", *Speech Communication*, 1996, Vol. 20, pp. 13-22.

7. V. Nagesha & L. Gillick, "Studies in Transformation-Based Adaptation", *Proc. IEEE Int Conf. Acoustics, Speech & Signal Processing, ICASSP-97,* Munich, Germany, April 1997, Vol. II, pp. 1031-1034.

COMPARING THREE METHODS TO CREATE MULTILINGUAL PHONE MODELS FOR VOCABULARY INDEPENDENT SPEECH RECOGNITION TASKS

Joachim Köhler

German National Research Center for Information Technology (GMD)
Institute for Media Communication (IMK),
53754 Sankt Augustin, Germany
Joachim.Koehler@gmd.de

ABSTRACT

This paper presents three different methods to develop multilingual phone models for flexible speech recognition tasks. The main goal of our investigations is to find multilingual speech units which work equally well in many languages. With this universal set it is possible to build speech recognition systems for a variety of languages. One advantage of this approach is to share acoustic-phonetic parameters in a HMM based speech recognition system. The multilingual approach starts with the phone set of six languages ending up with 232 language-dependent and context-independent phone models. Then, we developed three different methods to map the language-dependent models to a multilingual phone set. The first method is a direct mapping to the phone set of the International Phonetic Association (IPA). In the second approach we apply an automatic clustering algorithm for the phone models. The third method exploits the similarities of single mixture components of the language-dependent models. Like the first method the language specific models are mapped to the IPA inventory. In the second step an agglomerative clustering is performed on density level to find regions of similarities between the phone models of different languages. The experiments carried out with the SpeechDat(M) database show that the third method yields in almost the same recognition rate as with language-dependent models. However, using this method we observe a huge reduction of the number of densities in the multilingual system.

1. INTRODUCTION

Over the last years automatic speech recognition systems have reached a level of quality which allows the introduction of commercial products. However, a new problem has occurred: the language-dependency of current recognition technology. The phonetic models used in state-of-the-art systems are extremely language-dependent. The overall goal of our research activities is to create a multilingual and almost language independent recognition system which works in the most important languages of the world. We started our multilingual approach with OGI MLTS database [15] based on the work of [1]. Nowadays, even larger multilingual databases are available like SpeechDat(M)[1], Call-Home etc. These databases allow a robust modeling of phonetic units for different languages. Instead of using language-dependent acoustic models our approach tries to exploit the acoustic-phonetic similarities of sounds across languages. This approach has two main advantages. First, the number of HMM parameters can be reduced significantly if it is possible to share phone models in different languages. Second, these multilingual models speed up the process of cross-language transfer. With the multilingual phone models the huge data collection process can be avoided or at least it can be reduced. This paper shows different approaches to achieve the goal to exploit the acoustic-phonetic similarities.

The paper is organized as follows: First, we present three different methods to create multilingual phone models using HMM technology. Then we perform our experiments with a language-dependent system covering six languages. These multilingual experiments are then given in the following chapter. At the end we give a summary of the current research status and a perspective for future research activities.

2. MULTILINGUAL PHONE MODELING

This section shows different approaches to find multilingual phone models for automatic speech recognition tasks. One central problem is to detect and to exploit the acoustic-phonetic similarities across languages. Which sound in one language is similar enough to a sound of another language to provide only one common model? This question leads to the definition of a similarity measurement of speech sounds. The other question is, if the phone is the optimal entity to exploit the similarities. Or is another speech unit like a sub phone unit or a single density of a continuous density HMM (CDMM) more appropriate to create multilingual models. The overall goal of the different approaches to find multilingual speech units is to generate models which perform as well as language-dependent models for different recognition tasks. Thus, we can formulate the task to create accurate acoustic models which also exploit the similarities across languages.

2.1. Mapping to the IPA based phone set (IPA-MAP)

The most obvious approach is to map the language-dependent models to the appropriate phone of the inventory of the International Phonetic Association (IPA). Here, the phonetic mapping is performed with phonetic knowledge rather than with some statistical based similarity measurement. Most of the phonetic inventories which are in use are based on IPA, like SAMPA, WORLDBET, TIMITBET or SPICOS. The rule of the mapping of the language-dependent phones $Ph_{l,i}^{LDP}$ to the multilingual phone units is:

$$Ph_{l,i}^{LDP} \rightarrow Ph_j^{IPA} \qquad (1)$$

The mapping is performed for each language. All phonetic segmentation and transcription files (label files) are transformed to

[1] For information about SpeechDat see the following URL's:
http://www.phonetik.uni-muenchen.de/SpeechDat.html
http://www.icp.grenet.fr/ELRA/home.html

the IPA based inventory. After this mapping a Viterbi based HMM Maximum Likelihood training is performed. Figure 1 shows the different steps of the approach IPA-MAP.

```
loop over all languages
    loop over all phones of one language
        mapping of the language-dependent phones to IPA phone:
        $Ph_{l,i}^{LDP} \rightarrow Ph_j^{IPA}$
        add to mapping file
transformation of the label-files using the mapping file
HMM-training over all languages:
- HMM-init
- HMM-Viterbi training, 6 iterations
```

Figure 1: Algorithm IPA-MAP

The main advantage of this approach is the simple way of getting multilingual models. Further, the final IPA-based models have a clear representation in the multilingual context and the cross-language transfer is also very simple. The sounds of the new language can be extracted very easily from the multilingual phone library. On the other hand the direct use of IPA does not consider the spectral properties and the statistical similarities of the phone models. Further, the IPA-based units do not model some language-dependent properties of the sounds. This can yield in a decrease of the accuracy of the acoustic models. This problem will be more severe as more languages will be included in this approach. Another disadvantage is that some inconsistencies of different phone systems of different languages and inventories can hurt this method.

2.2. Multilingual Phone Clustering (MUL-CLUS)

In this approach the language-dependent phone models are mapped to a multilingual set using a bottom-up cluster algorithm. Therefore, a similarity between two phone models has to be defined. In this work we apply a log-likelihood LL based distance measure. The distance between two phone models λ_i and λ_j is:

$$D_{LL}(\lambda_i, \lambda_j) = LL_i^i - LL_j^i \quad (2)$$

$$D_{LL}(\lambda_i, \lambda_j) = \log p(X_i|\lambda_i) - \log p(X_i|\lambda_j) \quad (3)$$

where λ_i is the model of phone i. The data is given by the token X_i. Respectively, the distance $D_{LL}(\lambda_j, \lambda_i)$ is given by:

$$D_{LL}(\lambda_j, \lambda_i) = LL_j^j - LL_i^j \quad (4)$$

$$D_{LL}(\lambda_j, \lambda_i) = \log p(X_j|\lambda_j) - \log p(X_j|\lambda_i) \quad (5)$$

Because the distances are not symmetric we calculate the average distance:

$$D_{LL}(\lambda_i; \lambda_j) = \frac{1}{2}(D_{LL}(\lambda_i, \lambda_j) + D_{LL}(\lambda_j, \lambda_i)) \quad (6)$$

At each cluster step the most similar pair of clusters are merged to a new cluster. This means that the two clusters $\hat{C}_i$ and $\hat{C}_j$ of all cluster pairs C_i and C_j with the smallest distance are merged:

$$(\hat{C}_i, \hat{C}_j) = \underset{C_i, C_j}{\operatorname{argmin}} D(i, j) \quad (7)$$

Because the estimation of the new phone models of the merged cluster is difficult to achieve the distance is always computed with the original language-dependent models which are the basic elements of one cluster. Hence, the distance between two clusters are determined with the furthest neighbor criterion. Therefore, we calculate the maximum distance of the initial clusters C_k^0 and C_l^0 which are in this case the language-dependent phone units.

$$(\hat{C}_i, \hat{C}_j) = \underset{k \in C_i, l \in C_j}{\operatorname{argmax}} D(k, l) \quad (8)$$

The usage of the furthest neighbor criterion has also the advantage to avoid huge log-likelihood calculations. The calculation of equation 6 requires also the data of the phone models. The data corresponds to the phone tokens which are extracted from the phonetic label files. Each phone has a pool of tokens which are used for the distance calculation. The number of tokens of each language-dependent phone unit is set to 500.

The complete algorithm to create multilingual phone models using clustering methods is given in figure 2.

```
loop over all languages
    HMM Viterbi training
    create language-dependent phone models
init: define a set of initial clusters from language-dependent phones $C_i := \{Ph_i\}$
Compute a symmetric distance matrix
while ($D_{min} < D_{thres}$)
    find pair of clusters with the minimum distance $D_{i,j}^{min}$
    Merge the two clusters $C = C_i \cup C_j$
    update the distance matrix
mapping of the language-dependent phones to the multilingual clusters
HMM-training over all languages:
- HMM-init
- HMM-Viterbi training, 6 iterations
```

Figure 2: Algorithm to create multilingual phone models using phone distance measurement and clustering (MUL-CLUS)

The cluster process continues until all calculated cluster distances are higher than a pre defined distance threshold. Alternatively, the clustering stops if a specified number of final clusters is achieved. After the clustering is finished we can use the cluster information to map the language-dependent models to the multilingual inventory. All label files are processed with this mapping information. Then the HMM models are trained with the maximum likelihood based Viterbi training.

The automatic clustering has the advantage to use statistic measurement based on HMM technology which is also used during recognition. The disadvantage is that the final multilingual units lose some clear representation and it is more difficult to transfer this models to a new language.

2.3. IPA-based Density Clustering (IPA-OVL)

The previous two approaches try to create complete multilingual phone models. This means that all parameters (i.e. sub phone units, densities of a CDHMM) of one model are shared across the different languages. On the other hand there are several language specific properties of the sounds. They exist due to different phonetic context, speaking style and rate, prosodic features and allophonic variations. To cover these effects we have presented a novel approach to create multilingual phone models [15]. Instead of complete overlapping phone models we assume that there are language-independent realization. This approach is achieved by using mixture densities. Figure 3 shows the idea of this method. There are regions of one IPA sound which are used in one, two or three languages. In this example the nasal [m] occuring in the languages German, Spanish and English has mixture components which are used in one, two or all three languages.

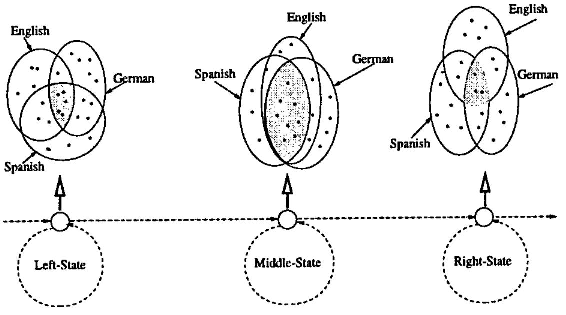


Figure 3: principles of the method IPA-OVL (two dimensional case).

The creation of the multilingual models is shown in figure 4. First, the language-dependent models are trained as before. Each language-dependent phone consists of 3 segments (sub phone units) each modeled by a mixture density. This is expressed by:

$$\lambda_{l,p}^{mono} = \left\{ S_{l,p,1}^{mono}, S_{l,p,2}^{mono}, S_{l,p,3}^{mono} \right\} \qquad (9)$$

where l is the language index and p in the phone index.

In the second step the mixtures of the language-dependent segments which belong to the same IPA-based phone are collected in one common pool of densities. Then we apply an hierarchical agglomerativ cluster algorithm to find and merge similar densities. The clustering is performed for each segment separately.

Because we work in our system with global variance values we use only the mean vectors for clustering. As distance measure giving the similarity between μ_i und μ_j the weighed L1-norm is applied:

$$D(\lambda_j; \lambda_i) = \frac{N_j N_i}{N_j + N_i} \sum_{d=1}^{D} |\mu_{i,d} - \mu_{j,d}| \qquad (10)$$

In previous investigations we found that is important to normalize the distance by the number of occurrences N_i und N_j which give information how often the densities are seen during training. This normalization avoids the generation of very big clusters which dominate the small clusters. One important aspect is that all clusters should have a similar number of elements. Otherwise the resulting clusters lose their power to discriminate between different sounds.

```
loop over all IPA-based phones
    loop over all three segments of one phone
        create a pool of densities belonging to the same IPA-based segment
        calculate the distance matrix for each pool of densities
minimum number of densities
    loop over all IPA-based phones
        loop over all phone segments (1,2,3)
            find pair of clusters $C_i$, $C_j$ with the minimum distance $D_{(i,j)}$
            merge the two clusters: $C = C_i \cup C_j$
            and remove cluster $C_i$ and $C_j$
            update distance matrix
```

Figure 4: Algorithm to create multilingual mixture densities (IPA-OVL)

For each pool of densities a distance matrix is calculated using equation 10. After each clustering step the overall number of densities is reduced by one element. The new cluster is given by the averaged mean vector of the two merged clusters. The clustering is finished if the complete system has a pre-defined number of densities. After finishing the cluster algorithm we have for each IPA-based phone a multilingual mixture density. Whereas the mixture density has multilingual regions the mixture weights are still language-dependent. For the calculation of the emission probabilities we use:

$$b_s(\vec{x}) = \sum_{m=1}^{M_s} c_{s,m}^{LDP} \mathcal{N}(\vec{x}, \vec{\mu}_{s,m}^{IPA}) \qquad (11)$$

Hence, this approach has some similarities to the semi-continuous HMMS. However, here the densities are shared only for one segment of one IPA-based phone across different languages. As final step the parameters of the multilingual mixture densities are reestimated during a Viterbi training. With this kind of multilingual modeling we also achieve a huge reduction of parameters in multilingual system. The combination of language-specific properties and automatically detection of multilingual realizations we exploit the acoustic-phonetic similarities in an optimal way.

3. EXPERIMENTS

In this section we perform several tests to compare the multilingual approaches. First, we describe briefly the speech engine. Second, we present the multilingual system using the language-dependent models. This system serves as comparison to the three previously described methods.

3.1. Description of the HMM-Based ASR system

For our investigations we use the SIEMENS HMM-based speech engine. The feature extraction generates every 10 ms a frame consisting of 24 mel-scaled cepstral, 12 Δ cepstral, 12 $\Delta\Delta$cepstral, 1 energy, 1 Δ energy and 1 $\Delta\Delta$ energy components. Each frame is processed by a LDA transformation reducing the 51 components to 24 values. To work in a multilingual environment one single LDA is calculated for all different languages. The acoustic models are based on Continuous Density HMMs (CDHMM) with Gaussian density functions. In our investigations we work only with context-independent models which consist of 3 sub-phone units (phone segments). Each segment is modeled by two states with tied emission probability.

3.2. Multilingual System with language-dependent models

The multilingual system covers the six languages American English, French, German, Italian, Portuguese and Spanish. The speech material is taken from the SpeechDat(M) and the Macrophone databases. Because all databases have only an orthographic transcription, all systems must be bootstrapped to generate an initial segmentation and label files. The bootstrapping was carried out with multilingual phone models based on the IPA-MAP method. The evaluation and tests were carried out on word and phone level. The word recognition rates are important for a final application and the phone recognition rates give some detail information about the acoustic modeling accuracy.

The training of the models is performed with the phonetic rich sentences of the databases. This should guarantee the vocabulary independence of the acoustic models. These models are also called *Type-In* models. The amount and structure of the training and test material is given in table 1. The training is performed with more than 4000 speakers and more than 35K sentences. The duration of the training material is almost 32 hours of pure speech without silence. The overall number of language-dependent phone units is 232. Italian has the greatest number of phones (49) because the SAMPA inventory distinguish between short and long consonants. Spanish has the smallest number using only 31 phones. The complete system has 31999 densities which means that in average each of the 232 language-dependent phone models have 45 densities.

After the training the models are tested on an isolated word and a phone recognition task. The recognition results for isolated words are summarized in table 2. The vocabulary size of this task varies between 47 and 70 words for the languages taken from SpeechDat(M). For American English the vocabulary size is 685 because there is no core test set for application words. The best results are achieved for German (96.6%). Also for the other 4 European languages we get results better than 90%. The result for American English is only 64.9% due to the high perplexity of the recognition task.

	#speaker	#utt.	hour.min	#
	tr-dev-te	Tr-Utt	Tr-Time	phones
French	667-166-167	6.0K	5.03	37
German	667-166-167	5.0K	4.18	38
Italian	667-166-167	5.8K	4.15	49
Portuguese	667-166-167	5.9K	7.33	38
Spanish	667-166-167	6.0K	5.38	31
Am.-English	1000-500-500	6.4K	5.12	39
Overall	4335-1330-1335	35.1K	31.59	232

Table 1: Structure of the training and test databases using SpeechDat(M) and Macrophone: tr ≙ number of speakers for training; dev ≙ number of speakers for developing purposes; te ≙ number of speakers for testing; Tr-Utt ≙ number of phonetic rich training sentences; Tr-Time ≙ time and duration of phonetic rich training sentences; number of phone units per each language

Language	#Rec-.	Voc.	*Rec.-*
	Tokens	Size	*Rate*
French	1420	57	92.2%
German	949	49	96.6%
Italian	983	47	94.4%
Portuguese	931	61	93.0%
Spanish	1242	70	93.3%
Am.-English	2612	685	64.9%
Average	–	–	89.0%

Table 2: Isolated word recognition rate for SpeechDat(M) and Macrophone database; Rec-Tokens: number of tested words; Voc. size: size of the vocabulary (perplexity); Rec. rate: word recognition rate

In the second test phone recognition rates are measured. The results given in table 3 including insertions, deletions and substitutions. For the continuous phone recognition task language-dependent bigram models are used to achieve a higher phone accuracy. It is very obvious that for Spanish and Italian the best phone recognition rates are achieved (56.9% and 53.2%). Both languages have a clear vowel structure. Also for German, French and Portuguese the recognition rates varies between 47.0% and 48.5%. Only for American English the recognition result ends

Language	#Rec-.	Voc.	*Phone*
	Tokens	Size	*Acc.*
French	12964	37	48.3%
German	12839	38	48.5%
Italian	10804	49	53.2%
Portuguese	21751	38	47.0%
Spanish	17512	31	56.9%
Am.-Englich	10815	39	37.7%
Average	–	–	48.6%

Table 3: Continuous phone recognition rate for SpeechDat(M) and Macrophone including deletions, insertions and substitutions

	LDP	IPA-MAP	MUL-CLUS	IPA-OVL
French	92.2%	90.9%	90.8%	92.5%
German	96.6%	91.6%	94.8%	96.5%
Italian	94.4%	93.6%	94.0%	93.7%
Portuguese	93.0%	89.6%	91.9%	91.9%
Spanish	93.3%	92.5%	93.3%	93.1%
Am.-English	64.9%	56.5%	57.0%	63.2%
Average	89.0%	85.5%	86.9%	88.5%

Table 4: isolated word recognition rates using the different multilingual approaches

with a disappointing 37.7% rate. One reason for this result could be the quality of the orthographic and phonetic transcription of the Macrophone database. In other investigation the results for American English are very similar to results in French or German [18].

Altogether the results on word and phone level show that it is possible to create task independent models with phonetic rich training material. These models are compared in the following section with the multilingual approaches.

3.3. Results using the Multilingual Approaches

Table 4 summarizes the isolated word recognition rates of the three different approaches in comparison to the language-dependent modeling. For these tests the number of densities was almost the same to achieve a fair comparison. The method IPA-OVL outperforms the other two methods (IPA-MAP and MULS-CLUS) and it was nearly as good as with the language-dependent models. The decrease in recognition rate was only 0.5% with only 13K densities instead of 31K densities in the language-dependent case. Hence, the method IPA-OVL is able to detect and exploit the acoustic-phonetic similarities across the phones of different languages. The data-driven phone clustering approach (MUL-CLUS) performs also better than the direct and simple mapping to the IPA inventory. For this two methods which model complete multilingual phones the decrease of recognition rate was 3.5% (IPA-MAP) and 2.1% (MUL-CLUS). Before we give a final conclusion the detailed results of the three methods are discussed.

IPA-MAP

The method IPA-MAP maps the 232 language-dependent models to 95 multilingual models. There are 13 phones (plosives, fricatives and nasals) which occur in all six languages. Table 5 gives an overview how many phones are used in different languages. This table also shows that 48 phones are still monolingual because they occur only in one language. However, the number of system parameters is drastically reduced. The number of densities decreases from 31999 to 13555 which reduces memory and computational resources of the multilingual recognition system significantly. However, the isolated word recognition rate decreases from 89% to 85.5%.

Whereas the decrease for the four Romance languages is small the reduction for German and American English is 5.0% and 8.4% respectively. Possible explanations for this effect are:

- differences in the quality and recording conditions of Macrophone and SpeechDat(M) databases:
  Although a channel compensation algorithm is used not all differences in the databases can be removed. This would at least explain the reduction of the American system.

# La.	# Ph.	list of phones
6	13	b d f g j k l m n p s t z
5	7	ʃ ɔ a r u v w
4	7	ɛ ŋ ɲ ʒ e i o
3	3	ə ʎ tʃ
2	17	ʊ œ ɽ ʀ ɑ ʏ ɛ̃ ɔ̃ ĩ aɪ aʊ dʒ h i: s: x θ
1	48	æ ç ɪ ø ø: β ɛ: ʃ: ɣ ð ɲ: ɔʏ ɝ ʎ: ɐ ɥ ʌ ɑ̃ ɐ̃ Ĩ õe a: b: d: dʒ: dz e: ei f: g: j: j̃ k: l: m: n: o: oʊ p: pf tʃ: t: ts u: ũ v: w̃ y ɔi

Table 5: Multilingual inventory using IPA-MAP

- sensitivity of the models for big vocabulary size:
  If the recognition task has a very high perplexity (in this case it is 685) very exact acoustic models are required. A small degradation of the models yields in a severe reduction of recognition rate.
- dominance of the Romance language in comparison to Germanic languages:
  Four of the six languages belong to the Romance language family. Hence, the multilingual models are dominated by the Romanian languages. This would explain the decrease of the German system.
- Inconsistency of the different phone inventories:
  Whereas for the Romance languages SAMPA is used, the German lexicon is based on SPICOS and the American lexicon uses TIMITBET. Although all inventories tries to realize the IPA-inventory there are some inconsistencies and problems during the mapping. For example in SPICOS the affricates [tS], [dZ], [pf] and [ts] are divided in two single phones. Also in the CMU-lexicon we observed some differences to the other inventories which could not be resolved easily. The central phone [ɐ] and the back vowel [ʌ] have the same phoneme symbol / ah /. Hence, the same symbol / ah / is used to transcribe the words "bottom" / b aa t **ah** m / and "cut" / k **ah** t /.

MUL-CLUS

The data-driven method MUL-CLUS yields in a higher recognition rate than the method IPA-MAP. Especially for German the results are much better. Instead of a reduction of 5.0% we observe only a decrease of 1.8%. However, the reduction for American English is still very obvious (7.9%). For this experiment the final number of multilingual phone units was chosen to 95 to have the same number of phones as before. The remaining clusters differs from the IPA-based mapping. The biggest cluster contains the fricatives [f], [s] of all six languages. Table 6 shows a selection of generated phone clusters. There are also some clusters which have same elements as with the IPA-MAP method. These clusters contain the nasals [m] and [n]. Phones which differ only in the phonetic length are very often mapped to the same cluster, especially for consonants. However, we also have 50 clusters with only one element. This means that we have still a huge number of monophones. Further, experiments were carried out with a varying size of final multilingual phone clusters. An observable decrease in recognition rate was observed when the 232 language-dependent models were clustered to less than 130 multilingual phones.

IPA-OVL

Here the clustering was performed on density level. The final

#CL	Cluster elements
15	f^{AE} f^{SP} f^{IT} f^{GE} f^{PT} f^{FR} $\mathrm{f{:}}^{IT}$ s^{AE} s^{GE} s^{PT} s^{FR} $\mathrm{s{:}}^{IT}$ s^{SP} s^{IT} θ^{SP}
12	p^{AE} p^{SP} p^{IT} p^{FR} p^{PT} p^{GE} t^{SP} t^{IT} t^{PT} t^{FR} t^{GE} $\mathrm{t{:}}^{IT}$
10	j^{AE} $\mathrm{i{:}}^{AE}$ i^{SP} i^{IT} i^{PT} i^{FR} $\mathrm{i{:}}^{GE}$ j^{SP} j^{IT} j^{PT}
7	m^{AE} m^{SP} m^{IT} m^{FR} m^{PT} m^{GE} $\mathrm{m{:}}^{IT}$
7	n^{AE} n^{SP} n^{IT} n^{GE} n^{FR} n^{PT} $\mathrm{n{:}}^{IT}$

Table 6: Selection of multilingual phone clusters generated with MUL-CLUS

number of densities was set to 13K. After the clustering process there were 7720 density clusters with more than one element (multilingual clusters) and 5280 monolingual clusters. This means that 25K of the 31K language-dependent densities are mapped to a multilingual cluster. The method IPA-OVL shows a significant improvement for the American system. The decrease was now only 1.7% in comparison to the language-dependent case.

4. SUMMARY AND CONCLUSION

In this paper we demonstrated the usefulness and feasibility of the multilingual approach. First, a telephone-based multilingual speech recognition system was built for 6 languages. The language-dependent phonetic models can be used for a vocabulary independent recognition tasks. Second, we developed and compared three different methods to create multilingual phone models. The best result was achieved with the method IPA-OVL which exploits the acoustic-phonetic similarities in an optimal way. However, this method works on the density level rather than on a complete phone level. Hence, it is important to consider the language-dependent properties of the phones even if they belong to the same IPA-based phone. The main advantage of the data-driven methods are the higher recognition rate and the fact that the final number of parameters can be adjusted during clustering. In all our investigations we used only context-independent models. Now it would be interesting to know how these methods would work with context-dependent models. Further, more languages of other language families should be integrated in this multilingual approach.

5. REFERENCES

[1] O. Andersen, P. Dalsgaard, W. Barry: *Data-Driven Identification of Poly- and Mono-phonemes for four European Languages. Proc. Eurospeech 1993*, 759–762, Berlin, 1993.

[2] O. Andersen, P. Dalsgaard: *Language-Identification Based on Cross-Language Acoustic Models and Optimised Information Combination. Proc. Eurospeech 1997*, 67–70, Rhodos, 1997.

[3] K.M. Berkling: *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters.* Oregon Graduate Institute of Science & Technology, 1996.

[4] U. Bub, J. Köhler, B. Imperl: *In-Service Adaptation of Multilingual Hidden-Markov-Models. Proc. ICASSP 1997*, 1451–1454, München, 1997.

[5] P. Bonaventura, F. Gallocchio und G. Micca: *Multilingual Speech Recognition for Flexible Vocabularies. Proc. Eurospeech 1997*, 355–358, Rhodos, 1997.

[6] G.L. Campbell: *Concise Compendium of the World's Languages.* Routledge, New York, 1995.

[7] C. Corredor-Ardoy, J. Gauvin, M. Adda-Decker, L. Lamel: *Language Identification with Language-Independent Acoustic Models. Proc. Eurospeech 1997*, 55–58, Rhodos, 1997.

[8] M. Falkhausen, H. Reininger, D. Wolf: *Calculation of Distance Measures Between Hidden Markov Models. Proc. Eurospeech 1995*, 1487–1490, Madrid, 1995.

[9] J.T. Foote, H.F. Silverman: *A Model Distance Measure for Talker Clustering and Identification. Proc. ICASSP 1994*, 317–320, Adelaide, 1994.

[10] J. Glass, et al.: *Multilingual Spoken Language Understanding in the MIT VOYAGER System.* Speech Communication, vol. 17, 1–18, 1995.

[11] A. Hauenstein, E. Marschall: *Methods for Improved Speech Recognition Over the Telephone Lines. Proc. ICASSP 1995*, 425–428, Detroit, 1995.

[12] J.L. Hieronymus: *ASCII Phonetic Symbols for the World's Languages: Worldbet.* Bell Labs Technical Memorandum, 1993.

[13] International Phonetic Association: *The International Phonetic Association (revised to 1993) – IPA chart. Journal of the International Phonetic Association*, vol. 1, Nr. 23, 1993.

[14] B.H. Juang, L.R. Rabiner: *A probabilistic distance measure for hidden Markov models.* Bell Syst. Tech. J., vol. 64, Nr. 2, 391–408, 1985.

[15] J. Köhler: *Multi-Lingual Phoneme Recogntion Exploiting Acoustic-Phonetic Similarities of Sounds. Proc. ICSLP 1996*, 2195–2198, Philadelphia, 1996.

[16] J. Köhler, : *Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks. Proc. ICASSP 1998*, 417–420, Seattle, 1998.

[17] P. Ladefoged, I. Maddieson: *The Sounds of the World's Languages.* Blackwell Publishers, Oxford, 1995.

[18] L.F. Lamel, J.L. Gauvain: *Cross-Lingual Experiments with Phone Recognition. Proc. ICASSP 1993*, 507–501, 1993.

[19] L.F. Lamel, M. Adda-Decker, J.L. Gauvin: *Issues in Large Vocabulary, Multilingual Speech Recognition. Proc. Eurospeech 1995*, 185–188, Madrid, 1995.

[20] Y.K. Muthusamy, A. Cole, B.T. Oshika: *The OGI Multi-Language Telephone Speech Corpus. Proc. ICSLP 1992*, 895–898, Banff, 1992.

[21] A. Sankar, F. Beaufays, V. Digalakis: *Training Data Clustering For Improved Speech Recognition. Proc. Eurospeech 1995*, 503–506, Madrid, 1995.

[22] T. Schultz, A. Waibel: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets. Proc. Eurospeech 1997*, 371-374, Rhodos, 1997.

Language adaptive LVCSR through Polyphone Decision Tree Specialization

T. Schultz[1] *and A. Waibel*[2]
{tanja@ira.uka.de}

[1] Interactive Systems Laboratories
University of Karlsruhe
Aussfarengarten 5a
76131 Karlsruhe, Germany

[2] Universität Karlsruhe, ILKD
Am Fasenengarten 5
D-76128 Karlsruhe, Germany

ABSTRACT

With the distribution of speech technology products all over the world, the fast and efficient portability to new target languages becomes a practical concern. In this paper we explore the relative effectiveness of porting multilingual recognition systems to new target languages with very limited adaptation data. For this purpose we introduce a polyphone decision tree specialization method. Several recognition results are presented based on mono- and multilingual recognizers developed in the framework of the project GlobalPhone which investigates LVCSR systems in 15 languages.

1. Introduction

With the distribution of speech technology products all over the world, the fast and efficient portability to new target languages becomes a practical concern. So far one major time and costs limitation in developing LVCSR systems in new languages is the need of large training data. According to the amount of data used for porting acoustic models to a new target language we differentiate three aspects of research:

- ⋆ Cross-language transfer (no data)
- ⋆ Bootstrapping (much data)
- ⋆ Language adaptation (very limited data)

The term *cross-language transfer* refers to the technique where a system developed in one language (group) is applied to recognize another language without using any training data of the new language. We do not distinguish whether the transfer to the target language is done from one language or from a group of languages. Research focuses on the questions whether cross-language transfer from one language to another language of the same family performs better than across family borders [4], and second if the number of languages used for training the transfer models influences the performance on the target language [7], [13]. Results seems to indicate a relation between language similarity and cross-language performance [4], [3]. Furthermore it is clearly shown that multilingual transfer models outperform monolingual ones [3], [14].

The key idea in the *bootstrapping* approach is to initialize a recognizer in the target language by using already developed acoustic models from other language(s) as seed models. After the initialization step the resulting system is completely rebuild using large training data of the target language. This idea was first proposed by Zue and evaluated by [6] and [15] showing that crosslanguage seed models perform better than flat starts or random models. Recently the usefulness of multilingual phonemic inventories and multilingual phoneme models as seed models have been demonstrated by [9], [11].

Language	Abbr	Utts	Spks	Units	Hours
Ch-Mandarin	CH	8529	112	219K	26.7
Croatian	CR	3374	72	89K	12.0
English (WSJ)	EN	7137	83	129K	15.0
French (Bref)	FR	7143	74	123K	13.9
German	GE	9173	71	132K	16.7
Japanese	JA	9096	108	212K	22.9
Korean	KO	6335	80	301K	16.4
Spanish	SP	5419	82	138K	17.6
Turkish	TU	5466	79	87K	13.2
Total		68276	839	1554K	170.4

Table 1: GlobalPhone database used for experiments

The *language adaptation technique* lies between the two extremes in terms of available training data. In this approach an existing recognizer is adapted to the new target language with only very limited data. [15],[9], [10] focus on two issues: first the amount of data needed to get reasonable results, second the question of finding suitable acoustic models to start from. For the first question they found -coincident to our expectation- that the language adaption performance is strongly related to the amount of data used for adaptation. [15] proved that the number of different speakers used for training is more critical than the number of utterances. The question of suitable models to start from was investigated by [9] and [10] comparing the effectiveness of multilingual acoustic models. Again it could be shown that multilingual models outperform monolingual ones.

Previous systems which combined multilingual acoustic models have been limited to small tasks and context independent modeling. Since for the monolingual case the use of larger phonetic context windows has proven to increase the recognition performance significantly, such improvements extend naturally to the multilingual setting. The idea how to construct context dependent multilingual models was first proposed by [5] and [14]. For the language adaptation purpose we intend to exploit the context information learned from several languages. How this information can be incorporated into the language adaptive process is still an open issue. In this paper we present a new approach to adapt polyphone decision trees to the new target language.

2. Multiple Languages

For our experiments we developed monolingual LVCSR systems in nine languages which will be introduced in this section. For training and testing we are using our multilingual database GlobalPhone.

2.1. The GlobalPhone Database

GlobalPhone currently consists of the languages Arabic, Chinese (Mandarin and Shanghai dialects), Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. In each of these languages we collected about 15 hours of speech spoken by 100 native speakers per language. Every speaker read several articles from a national newspaper. The articles were chosen from the areas: national politics, international politics, and economy. The speech data was recorded at a sampling rate of 48kHz using a close-talking microphone connected to a DAT-recorder. After transferring the sound data from DAT to hard disc it was downsampled to 16kHz, 16-bit. The GlobalPhone corpus is fully transcribed, and during validation process special markers were added for spontaneous effects like false starts, and hesitations. Further details about the GlobalPhone project are given in [12].

Since English and French are already available in very similar frameworks we decided not to collect additional data in these well covered languages but add the two databases Wall Street Journal (WSJ0, distributed by LDC) for English and Bref (BREF-Polyglot sub-corpus, distributed by Elsnet) for French to our training data. The resulting database covers 9 of the 12 most widespread languages of the world.

Throughout the experiments 80% of the speakers were used for training the acoustic models, 10% were defined as a test set, and the remaining 10% were kept as further cross-validation set. See table 1 for an overview of the database used throughout the experiments.

2.2. Monolingual Baseline Recognizers

We developed equally designed monolingual LVCSR systems in nine languages using our Janus Recognition Toolkit (JRTk). For each language the resulting baseline recognizer consists of fully continuous 3-state HMM systems with 3000 polyphone models. Each HMM-state is modeled by one codebook which contains a mixture of 32 Gaussian distributions. The preprocessing is based on 13 Mel-scale cepstral coefficients with first and second order derivatives, power and zero crossing rate. After cepstral mean subtraction a linear discriminant analysis reduces the input to 32 dimensions.

Language	Word based			Phoneme based		
	ER	Vocab	PP	ER	Vocab	PP
Ch-Mandarin	14.5	45K	207	45.2	141	12.5
Croatian	20.0	15K	280	36.7	32	9.6
English	14.0	64K	150	46.4	46	9.2
French	18.0	30K	240	36.1	38	12.1
German	11.8	61K	200	44.5	43	9.0
Japanese	10.0	22K	230	33.8	33	7.9
Korean	31.0	64K	130	36.1	43	9.9
Spanish	20.0	15K	245	43.5	42	8.2
Turkish	16.9	15K	280	44.1	31	8.5

Table 2: Word and phoneme based error rates (ER), vocabulary size, and trigram perplexity (PP) for nine languages

In table 2 we arranged the error rates[1], vocabulary size and trigram perplexities for the monolingual recognizer. Since the engines are the same across the languages, differences in the recognition performance are due either to language specific inherent difficulties or to differences in quality and quantity of the used knowledge sources and data. In our opinion it is misleading to infer from the given word error rates to language difficulties. On the one hand the concept of a word does not hold for each language (Chinese, Japanese, and Korean). On the other hand the word error rates are strongly affected by available corpus data and resulting artifacts like different vocabulary sizes, OOV-rates, language model perplexities, and last but not least by the human language expertise, which in our case is not comparable in all languages.

A reliable measure of the acoustic difficulties of the nine languages is the phoneme based recognition rate using a phoneme recognizer without any (phoneme) language model constraints. The results in table 2 indicate significantly differences in acoustic confusability between languages, ranging from 33.8% to 46.4% phoneme error rate. English seems to be the most hardest task in acoustical sense whereas Japanese is the easiest.

3. Multilingual Systems

In this section we describe our approach to create a multilingual recognizer engine by combining context dependent acoustic models across languages.

3.1. Global Phonetic Inventory

We intend to share acoustic models of similar sounds across languages for the adaptation purpose. Those similarities can be either derived from international phonemic inventories documented in Sampa, Worldbet, and IPA or by data-driven methods as proposed for example by [1].

In our work we defined a *global phoneme set* based on the phonemic inventory of the monolingual systems. Sounds

[1] Mandarin is given in character based error rate, Japanese in hiragana based error rate, and Korean in syllable based error rate

which are represented by the same IPA symbol share one common phoneme category. In case of five languages we started with 171 language specific phonemes and pooled them together into 85 phoneme categories. In case of nine languages we pooled 339 language dependent phonemes into 140 phoneme categories. Thus the phone-set compression rate of 49% in the five-lingual case increases to 41% in the nine-lingual case.

3.2. Multilingual acoustic model Combination

Based on the above described phoneme categories we designed multilingual systems by combining the language dependent acoustic models of the languages Croatian, Japanese, Korean, Spanish, and Turkish in two different ways and compared their effectiveness for the language porting purpose.

In system *ML-mix* we share all models across these five languages without preserving any language information. We build context dependent models by applying a decision tree clustering procedure which uses a question set of linguistic motivated phonetic context questions. We train the models by sharing the data of the five languages. In the second system *ML-tag* the phoneme model sharing across languages is performed by attaching a language tag to each of the phoneme categories in order to preserve the information about the language. The above described clustering procedure is enhanced by introducing questions about the language and language groups to which a phoneme belongs. Therefore the decision if phonetic context information is more important than language information becomes data-driven (see [14] for details).

We explore the usefulness of the two different modeling approaches by running three experiments on 7 recognizers summarized in table 3:

1. Monolingual baseline test: all five monolingual recognizers are tested on the corresponding language
2. Multilingual test: *ML-mix* and *ML-tag* are applied to recognize one of the five languages involved in training the multilingual models
3. Porting test: the five monolingual systems as well as *ML-mix* and *ML-tag* are applied to recognize German utterances.

The results of the multilingual test show that *ML-tag* outperforms the mixed system *ML-mix* by 5.3% (3.1% - 8.7%) error rate. This indicates that preserving the language information achieves better results with respect to the ideal situation that sufficient training data is available to build a language specific system. This finding is coincident to other studies [5], [9]. The porting test prove that *ML-mix* outperforms *ML-tag* in both techniques. This is evident since sharing information across languages augments the language robustness of the transfer system (see [14] and [10] for details).

Language	Mono	ML-tag	ML-mix
Croatian	26.9	31.9	35.0
Japanese	13.0	15.0	20.0
Korean	47.3	49.0	
Spanish	27.6	32.4	37.0
Turkish	20.1	21.3	29.0
Technique	Porting to German		
Crosslanguage	49.5-65.0	50.0	41.5
Bootstrap	28.4-50.5	35.7	29.2

Table 3: Word error rates of *ML-mix* versus *ML-tag*

3.3. Dictionary Mapping

For all our experiments we presume that a pronunciation dictionary for the target language is given in an arbitrary phoneme set. Since we are interested in time and cost effective algorithms we created dictionaries which are not already available from scratch by grapheme-to-phoneme tools. However we post-edit the dictionaries by human experts who added pronunciation variants and treated special events like acronyms.

Nevertheless for recognizing the target language with the *ML-mix* or *ML-tag* system we need to define an appropriate mapping from our global phoneme set to the target phonemes. We investigate two approaches to find this mapping: In the first approach we apply an heuristic IPA-based mapping, meaning that a human experts defines for each target phoneme the corresponding counterpart according to our IPA phoneme categories. In the second approach we perform a data-driven mapping by calculating a phoneme confusion matrix, and picking the phoneme as a counterpart which leads to the highest confusion with the target phoneme. For this experiment we assume that an accurate phoneme recognizer in the target language is already given. We calculated phonetic alignments of 500 utterances spoken in the target language and did a frame-wise comparison with the viterbi decoded alignment of the same 500 utterances using a multilingual recognizer. Our experiments show that the IPA-based approach outperforms the data-driven approach by 27.1% vs 34.3% word error rate for the bootstrap technique and 66.7% vs 74.5% word error rate for the cross-language transfer technique (see [13] for details).

4. Polyphone Decision Tree Specialization

When creating the *ML-mix* system we uses a divisive clustering algorithm that builds context querying decision trees [8]. As selection measure for dividing a cluster into two subclusters we used the maximum entropy gain on the mixture weight distributions. This clustering approach gave significant improvements across different tasks and languages [8]. Figure 1 shows for 10 languages the number of different models we can get when using different context sizes. As can be seen these numbers differs very much between the languages. These differences are due to the perplexity of the language, to the number of words in the training corpus, and to the length of the modeled word units. The latter is according to a contraint imposed by the decoder which limits the maximum context width to all phonemes within a word and up to one phoneme into the neighboring words. For example the extremely shortness of Korean units used in our recognizer results in zero polyphones of context larger than 2. While for Chinese, Japanese and Spanish the most frequent word length in the training data is 2 phonemes, it is 5 for Turkish and 6 for Russian. The most frequent numbers of phonemes in the dictionary various from 2 for Spanish to 9 for Turkish.

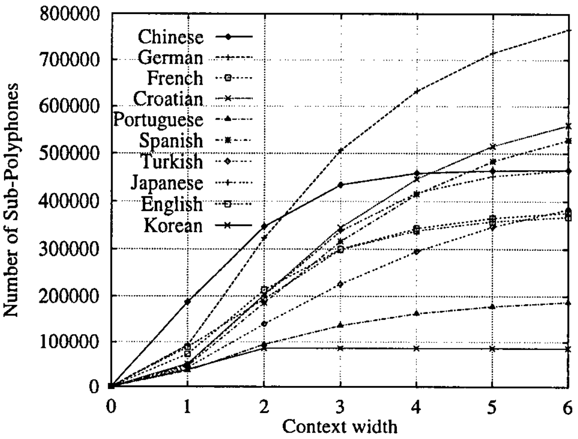


Figure 1: Different Sub-polyphones in training corpus

The concept of the IPA-based phoneme categories allows us to share context models across languages. To estimate the percentage of polyphone overlap between languages we define the non symmetric polyphone coverage measure as the number of polyphone occurrences in one language covered by polyphones in another language. In table 4 we give the triphone coverage for 10 languages. Here we distinguish between the coverage of polyphone types (upper row) and the coverage of polyphone occurrences (lower row), where the first one focus on the aspect whether common polyphones exists across languages, and the latter one focus on the aspect that frequent polyphones are more important to cover than rare ones. For example 33.6% of Japanese triphone occurrences are covered by German triphones, whereby 22.3% of the polyphone types are responsible for this coverage rate. On the other hand only 19.5% of German triphone occurrences are covered by Japanese polyphones. This effect is due to the Japanese phonotactic which only allows consonant vowel combinations.

From table 4 it is obvious that we should be aware of a large mismatch between represented polyphones in the multilingual decision tree and the observed polyphones in a new target language. We therefor specialize the already existing multilingual polyphone decision tree to the new language by continuing growing the decision tree. The limited amount of adaptation data is used to train separate mixture weight distribution for the resulting leaf nodes.

B/C	CH	DE	EN	FR	JA	KO	KR	PO	SP	TU
CH	100	0.1 5.3	0.3 6.8	0.1 5.8	0.1 4.2	0.0 5.3	0.1 4.2	0.1 5.4	0.1 5.3	0.2 4.9
DE	0.1 3.9	100	5.5 19.6	19.8 41.6	9.3 19.5	7.2 18.2	18.6 34.9	13.6 28.0	12.9 28.3	12.9 26.1
EN	0.6 5.2	5.4 18.1	100	6.5 18.6	1.8 8.9	3.4 11.6	1.5 7.7	0.9 6.6	1.3 6.6	3.8 9.2
FR	0.1 3.9	29.0 53.3	9.7 16.4	100	10.2 22.7	11.2 28.7	25.8 45.5	18.4 36.4	17.4 41.3	23.1 35.6
JA	0.2 2.5	22.3 33.6	4.5 9.9	16.8 37.4	100	9.8 25.6	16.0 29.2	11.0 27.6	13.6 31.2	25.9 52.5
KO	0.1 4.1	10.3 36.3	4.9 16.1	10.9 35.0	5.8 24.9	100	10.2 38.6	8.0 30.8	9.3 38.4	9.1 26.1
KR	0.2 1.8	39.0 68.8	3.2 5.0	37.0 64.7	14.0 28.2	15.0 34.5	100	31.0 63.0	34.3 61.8	31.5 50.4
PO	0.4 2.3	30.2 57.9	2.0 4.6	28.0 49.5	10.2 26.7	12.5 37.5	32.9 62.5	100	33.5 57.5	19.8 39.9
SP	0.2 2.5	25.4 60.2	2.7 5.6	23.5 60.1	11.2 34.0	12.9 40.1	32.2 64.2	29.7 58.2	100	17.5 41.0
TU	0.8 5.4	29.6 46.0	8.9 18.3	36.3 52.0	24.8 46.1	14.6 33.0	34.4 50.1	20.4 38.6	20.3 39.6	100

Table 4: Triphone Coverage for 10 languages

5. Language Adaptation to Portuguese

In the previous sections we report on the usefulness of multilingual acoustic model combination with respect to porting these acoustic models to the German language with the cross-language transfer and bootstrap technique. Now we investigate the benefit of these multilingual models in combination with the polyphone decision tree specialization (PDTS) for language adaption. We intend to adapt the different described multilingual systems to Portuguese. For adaptation we presume that a Portuguese dictionary as well as the recordings and transcriptions of 200 spoken utterances are given. Although [15] found that the number of speakers for adaptation is more critical than the number of utterances we decide to use 200 utterances spoken by only 7 different Portuguese speaker since at least in our dictation task it is more expensive to get single utterances of many different speakers than to get many utterances spoken by one speaker. The 200 utterances result in 25 minutes speech with 3370 spoken word units for adapting the acoustic models. The dictionary mapping was done according to our heuristic IPA-based mapping approach.

A subset of 96 uniformally selected utterances from 3 test speakers was used to carry out our experiments. The test dictionary has 7300 entries, the OOV-rate is set to 0.5% by including the most common words of the test set into the dictionary. A trigram language model with Kneser/Ney backoff scheme was calculated on 10 million word text corpus from Agency France Press interpolated with the GlobalPhone data leading to a trigram perplexity of 297.

5.1. Polyphone Coverage

Before applying our polyphone decision tree specializing approach we want to examine how well the 49 Portuguese monophones and resulting polyphones are covered by the nine- and five-language pool. Therefore we calculated the coverage of Portuguese polyphones according to our IPA phoneme categories. This measure indicates how well a not specialized polyphone decision tree fits to the target language. The coverage is shown in figure 2 for context width 0 (monophones) and 1 (triphones). The calculation of plotted coverage proceeds as follows: first we select that language among all pool languages which achieves the highest coverage for Portuguese. We then remove this language from the pool and calculate the coverage between Portuguese and each language pair resulting from the combination of removed language plus remaining pool language. The procedure is repeated for triples and so forth. Thus in each step we find the language which maximally complements the polyphone set.

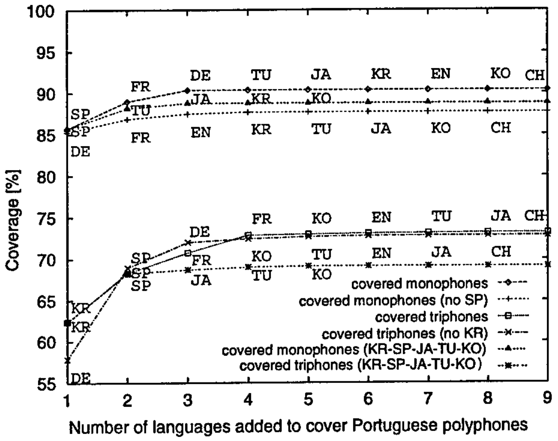


Figure 2: Portuguese polyphone coverage by nine languages

From the figure 2 we observed that as expected the coverage dramatically decrease for larger context (for quintphones a maximal coverage of 46% could be attained). After incorporating three languages the coverage of Portuguese monophones can not increased any further, limited to 91% with the nine language pool and dropping to 85% when the most important language for monophone coverage (SP) is removed from the language pool. The contribution of the Spanish phoneme set to the monophone coverage can not be compensate by other languages remaining in the pool. Second we found that when increasing the context width to 1 the coverage saturate after four languages. When increasing to context width to 2 we observed that at least five languages contribute to the quintphone coverage rate. Therefor we infer that increasing the context width requires more languages. For the context width 1 the main contribution comes from the Croatian language. Removing this language from the pool is nearly completely compensate by German and Spanish triphones. This indicate that Croatian, German, and Spanish polyphones covers a similar portion of the Portuguese triphones set. Whereas the curve (KR-SP-JA-TU-KO) indicates that the French language contribute unique polyphones which can not be recruited from other languages. In this case the lacking phonemes belong to the categories of nasal vowels. We conclude from this observation that when designing a language pool for adaptation purposes it is more critical to find a complement set of languages than to cover a large number of languages. Our method of calculating the polyphone coverage across languages can help to find such a complementary language set. From analyzing the polyphone coverage we draw the conclusion that using a polyphone tree even based on several languages can not be applied successful to Portuguese without adapting to the new contexts.

System	Data	Labels	Technique	Ptree
	Cross-language transfer			
S1	0	-	-	ML
S2	0	-	-	CI
	Language adaptation			
S4	100	initial	MLAdapt	CI
S5	100	initial	Viterbi	ML
S6	100	initial	MLAdapt	ML
S7	100	good	MLAdapt	ML
S8	200	good	MLAdapt	ML
S9	200	good	PDTS	ML-PO
	Bootstrap			
S3	100	initial	Rebuild	PO
S10	6600	good	Rebuild	PO

Table 5: Description of systems adapted to Portuguese

5.2. Results

Table 5 describes the systems used for our adaptation experiments, their performance on Portuguese is compared in figure 3. The column **Data** in table 5 refers to the number of recordings used as adaptation data. Applying no data results in a cross-language transfer approach as performed in the systems S1 and S2. Whereas the training based on 6600 utterances (S10) represents the bootstrap technique. For the systems S3 to S9 we used very limited data of 100 and 200 utterances.

Labels explains whether the phonetic transcription of the recordings are created based on the multilingual recognition engine *ML-mix* (Labels = initial) or based on good phonetic alignments which we presume to be already given (Labels = good). The latter was used to accelerate our adaptation process. In future work we will examine if we can get close to this label quality by iterating our adaptation approach.

The term **Technique** is related to the training approach applied to the systems. Viterbi refers to one iteration of viterbi training along the given labels. MLAdapt means Maximum Likelihood Adaptation technique, Rebuild refers to the iterative procedure of writing labels, viterbi training, model clustering, training, and writing improved labels. PDTS is the described Polyphone Decision Tree Specialization.

The **Ptree** item describes the origin of the polyphone decision trees. CI refers to context independent modeling, meaning that no polyphone tree is used, ML is the 3000 polyphone tree of system *ML-mix* and PO is a polyphone tree build exclusively on Portuguese polyphones. ML-PO refers to the regrown *ML-mix* polyphone tree applying PDTS.

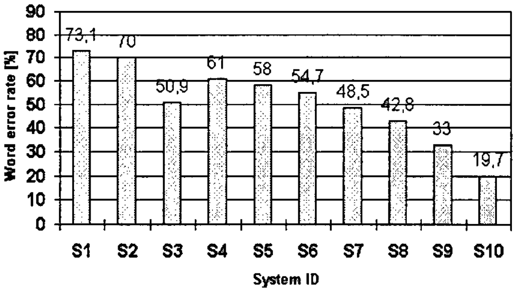


Figure 3: Language adaptation to Portuguese

As expected the recognition of Portuguese speech by running the five-lingual recognizer *ML-mix* without any training data results in extremely high word error rates of 73.1% for the context dependent system (S1) and slightly better error rates of 70% for the context independent system (S2). Therefor the initial labels are written with the multilingual context independent system S2. Using 100 of these initial labels for adapting the context independent multilingual system (S4) and the context dependent system by MLA (S6) or viterbi training (S5) shows a significant gain. In S3 the initial labels are used to completely rebuild a Portuguese system after bootstrapping from multilingual seed models. The comparison of S6 and S3 indicate that the adaptation of non matching polyphone trees is outperformed by the bootstrap technique (S3) even if data are very limited. Nevertheless the word error rate of the winning system S3 achieving 50.9% is still unsatisfying.

We obtain the next performance boost from using improved labels (S7) and double amount of adaptation data (S8). Finally we applied our PDTS approach (S9) which leads to significant improvements achieving 33% word error rate. This performance compares to 19.7% word error rate (S10) resulting from bootstrapping and rebuilding a Portuguese LVCSR system using 16 hours of speech spoken by 78 speakers. To summarize we get the highest performance gain in language adaptation from the PDTS technique, enlarging adaptation data, and improved labels, in this order.

6. Conclusion

In our language adaptive approach we explore the relative effectiveness of multilingual context dependent acoustic models in combination with a polyphone decision tree specialization (PDTS). We examine the profit when porting a multilingual engine to new target languages with very limited training data. The results are very promising achieving 33% word error rate for an Portuguese LVSCR system when using only 200 spoken utterances for adaptation.

7. Acknowledgment

The authors gratefully acknowledge all members of the GlobalPhone team for their great enthusiasm. We also wish to thank the members of the Interactive Systems Laboratories, especially Roald Wolff for his active support, great encouragement and contribution to this research.

References

1. O. Andersen et al.: *Data-Driven identification of Poly- and Mono-phonemes for four European Languages* in: Proc. Eurospeech, pp. 759-762, Berlin 1993.
2. J. Barnett et al.: *Multilingual Speech Recognition at Dragon Systems* in Proc. ICSLP, pp. 2191-2194, Philadelphia 1996.
3. U. Bub et al.: *In-Service Adaptation of Multilingual Hidden-Markov-Models*, Proc. ICASSP, pp. 1451-1454, Munich 1997.
4. A. Constantinescu et al.: *On Cross-Language Experiments and Data-Driven Units for ALISP* in: Proc. ASRU, pp. 606-613, St. Barbara, CA 1997.
5. P. Cohen et al.: *Towards a Universal Speech Recognizer for Multiple Languages* in: Proc. ASRU, pp. 591-598, St. Barbara CA, 1997.
6. J. Glass et al.: *Multi-lingual Spoken Language Understanding in the MIT Voyager System* in: Speech Communication (17), pp. 1-18, 1995.
7. S. Gokcen et al.: *A Multilingual Phoneme and Model Set: Towards a universal base for Automatic Speech Recognition* in: Proc. ASRU, pp. 599-603, St. Barbara, CA 1997.
8. M. Finke et al.: *Wide Context Acoustic Modeling in Read vs. Spontaneous Speech* in: Proc. ICASSP, Munich 1997.
9. J. Köhler: *Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks*, Proc. ICASSP, pp. 417-420, Seattle, 1998.
10. T. Schultz et al.: *Language independent and language adaptive LVCSR* in: Proc. ICSLP, pp. 1819-1822, Sydney 1998.
11. T. Schultz et al.: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets* in: Proc. Eurospeech, pp. 371-374, Rhodes 1997.
12. T. Schultz et al.: *The GlobalPhone Project: Multilingual LVCSR with Janus-3* in: Proc. SQEL, pp. 20-27, Plzeň 1997.
13. T. Schultz et al.: *Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages* in: Proc. Specom, pp. 207-210, St. Petersburg, Russia 1998.
14. T. Schultz et al.: *Multilingual and Crosslingual Speech Recognition* in: Proc. DARPA Workshop on Broadcast News Transcription and Understanding, Lansdowne, VA 1998.
15. B. Wheatley et al.: *An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language* in: Proc. ICASSP, pp. 237-240, Adelaide 1994.

Report of the plenary discussion on “Cross Language”

Chairperson: Timothy Anderson (WPAFB, USA)
Reporter: Johan Koolwaaij (KUN, the Netherlands)

Questions to *Adda-Decker*: Most questions were about the degree of realism in Adda-Decker’s view on interoperability in the future. *Van Leeuwen* asked if real time porting of speech recognition means that she expects rule based porting of language models in the future. *Adda-Decker* replied that it is more about general availability of the language models, for example via Internet. So real time porting means real time acquisition in this case.

Koehler asked how realistic is it to expect standardization of phone sets? According to *Adda-Decker* it is better to have something global you agree upon, than having detailed phone sets and no agreement.

Geoffrois inquired if automatic learning is really possible? *Adda-Decker* replied that this is only feasible for low complex applications, it is not to be expected for higher complex applications.

Hunt asked if *Micca’s* approach on ‘Multilingual Vocabularies in Automatic Speech Recognition’ is possible to apply within a language. *Micca* responded that he tried this already with reasonable success.

Anderson inquired why the performance of the *Ueblers* English recognizer is so poor compared to the other languages? *Uebler* remarked that it is difficult to compare recognizers because of differences in application, vocabulary size, language model, etc...

Eklund noted that speakers tend to use the native approach, but what happens if they don’t? Does it happen that they phones from a 3rd language, like when Swedish have to pronounce Aachen and don’t know any German, they sometimes choose an English pronunciation. *Uebler*: Not really looked into, but might very will be.

Cole to Hunt: What exactly is unsupervised adaptation in your ‘Military Operational Automatic Interpreting System’? *Hunt*: Adaptation without the system having any a priori knowledge about what the speaker is going to say.

Boves. How do you handle confirmation? *Hunt*: Confirmation is usually by gestures, like nodding in case of agreement. To be sure that the system translates the correct sentence, audio verification mode does exist.

Junqua: Is the system available in other languages? *Hunt*: All Dragon-Dictate languages are possible as input language.

Anderson: Why is the performance worse in case of a female speaker? *Hunt*: It is not the SNR, not the speech rate, not the amount of variability, but the fundamental frequency. Worst performing female had a F_0 of 300 Hz. There is a proprietary algorithm now which solves this problem to some extent.

Anderson: Are there plans for two-way translation? *Hunt*: One and a half way already exists: one way full translation, other way translation of numerals only.

Van Compernolle to Koehler: on “Comparing Three Different Methods to Create Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks” “Is the LDA done after the IPA mapping?” *Koehler*: Yes.

Kienappel: Does it happen that all phones from one language end up in one cluster? *Koehler*: Yes, and this is indeed a problem.

Imperl to Schultz: Would you expect improvement by using more languages with your Language adaptive LVCSR? *Schultz*: Going from 5 to 8 languages does not make real difference. Carefully choosing 5 languages to obtain a phone set with highest coverage helps a bit more.

Adda-Decker: OOV rate per language should be mentioned in the table, since it is proven to have a large impact. *Schultz* is aware, also comparing error rates over languages does not really make sense because of different types of units in some languages.

General discussion

Phone models

What are the best phone models? Context dependent (CD) models or context independent (CI) models? The outcome of the discussion is that this is really application dependent. In general more specific CD models are better that global CI models (*Koehler and Micca*) But the choice for the contexts of the CD models should be made carefully. Hunt adds that in general the segmentation for pronunciation guessing is better with CI models. Martine Adda says that CD models include the coarticulation effects.

Porting

What is the biggest issue in porting? Lack of data is an important issue when you want to start from CI models to go to CD models because data driven approaches to select the right contexts need lots of data (*Van Compernolle*). *Hunt* adds that the SQALE projects showed that the choice of units in different language is not straightforward. For example, Japanese and Turkish have a concept totally different from English. *Schultz*: So language-modeling issues are really important for languages other than English. *Adda-Decker*: For German OOV control is needed, but that implies control on the difficulty of the language, and German also has the compounding problem. Perplexity only is a rough indication of the complexity of the problem. *Schultz* mentions the issue of transcriptions in the original language. So far she used 'romanized' transcription, but at some point in time original transcription will be needed.

Phone sets:

What is the right phone set? IPA, SAMPA, xenophones, ... Knowledge or data driven? This appeared to be an endless discussion. Geoffrois says that standardizing the phone set is comparable to standardizing the words in a language. Others prefer a working solution. *Boves* concludes the discussion by remarking that the optimal solution does not exist, and that we necessarily should agree on a sub-optimal solution.

Similarity measures

Should phonetic similarity measures in perceptual space or spectral space? Van Compernolle starts off with the remark that all these measures also include differences in for example recording conditions as we have seen in a number of papers which makes it extremely difficult to interpret these measures. *Koehler* says that using these measures for clustering of phones over languages sometimes results in the unwanted effect that all phones from one language end up in one cluster. His working solution was to use SpeechDat corpora. Others also have good experience with these corpora.

AUTOMATIC LANGUAGE IDENTIFICATION

Marc A. Zissman, Kay M. Berkling

Information Systems Technology Group
Lincoln Laboratory
Massachusetts Institute of Technology
244 Wood Street
Lexington, MA 02420-9185
e-mail: MAZ@SST.LL.MIT.EDU
e-mail: KAY@SST.LL.MIT.EDU

ABSTRACT

Automatic language identification is the process by which the language of a digitized speech utterance is recognized by a computer. In this paper, we will describe the set of available cues for language identification and discuss the different approaches to building working systems. This overview includes a range of historic approaches, contemporary systems that have been evaluated on standard databases, as well as promising future approaches. Comparative results are also reported.

1. INTRODUCTION

Automatic language identification is the process by which the language of a digitized speech utterance is recognized by a computer. It is one of several processes in which information is extracted automatically from a speech signal.

Language-ID (LID) applications fall into two main categories: preprocessing for machine systems and preprocessing for human listeners. Figure 1 shows a hotel lobby or international airport of the future that employs a multi-lingual voice-controlled travel information retrieval system. If no mode of input other than speech is used, then the system must be capable of determining the language of the speech commands either while it is recognizing the commands or before it has recognized the commands. Determining the language during recognition would require many speech recognizers (one for each language) running in parallel. Because tens or even hundreds of input languages would need to be supported, the cost of the required real-time hardware might prove prohibitive. Alternatively, a language-ID system could be run in advance of the speech recognizer. In this case, the language-ID system would quickly list the most likely languages of the speech commands, after which the few most appropriate language-dependent speech-recognition models could be loaded and run on the available hardware. A final language-ID determination would be made only after speech recognition was complete.

Figure 2 illustrates an example of the second category of LID applications—preprocessing for human listeners. In this case, LID is used to route an incoming telephone call to a human switchboard operator fluent in the corresponding language. Such scenarios are already occurring today: for example, AT&T offers a *Language Line* interpreter service to, among others, police departments handling emergency calls. When a caller to *Language Line* does not speak English, a human operator must attempt to route the call to an appropriate interpreter. Much of the process is trial and error (for example, recordings of greetings in various languages can be used) and can require connections to several human interpreters before the appropriate person is found. As reported by Muthusamy et al. [33], when callers to *Language Line* do not speak English, the delay in finding a suitable interpreter can be on the order of minutes, which could prove devastating in an emergency. Thus, a LID system that could quickly determine the most likely languages of the incoming speech might be used to reduce the time required to find an appropriate interpreter by one or two orders of magnitude.

2. LANGUAGE IDENTIFICATION CUES

Humans and machines can use a variety of cues to distinguish one language from another. The reader is referred to the linguistics literature (e.g., [5, 6, 12]) for in-depth discussions of how specific languages differ from one another and to Muthusamy et al. [35], who has measured how well humans can perform language ID. In summary, the following characteristics differ from language to language:

- Phonology. A "phoneme" is an underlying mental representation of a phonological unit in a language. For example, the eight phonemes that comprise the word "celebrate" are /`s eh l ix b r ey t`/. A "phone" is a realization of an acoustic-phonetic unit or segment. It is the actual sound

THIS WORK WAS SPONSORED BY THE DEPARTMENT OF DEFENSE UNDER AIR FORCE CONTRACT F19628-95-C-0002. OPINIONS, INTERPRETATIONS, CONCLUSIONS, AND RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND ARE NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

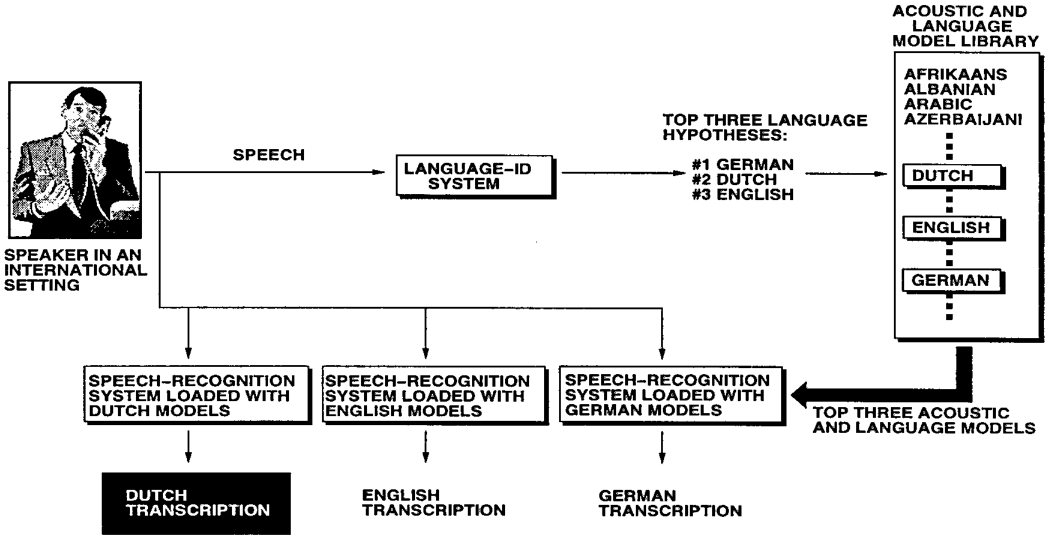

ACOUSTIC AND LANGUAGE MODEL LIBRARY
AFRIKAANS
ALBANIAN
ARABIC
AZERBAIJANI
DUTCH
ENGLISH
GERMAN
SPEAKER IN AN INTERNATIONAL SETTING
SPEECH
LANGUAGE-ID SYSTEM
TOP THREE LANGUAGE HYPOTHESES:
#1 GERMAN
#2 DUTCH
#3 ENGLISH
SPEECH-RECOGNITION SYSTEM LOADED WITH DUTCH MODELS
SPEECH-RECOGNITION SYSTEM LOADED WITH ENGLISH MODELS
SPEECH-RECOGNITION SYSTEM LOADED WITH GERMAN MODELS
TOP THREE ACOUSTIC AND LANGUAGE MODELS
DUTCH TRANSCRIPTION
ENGLISH TRANSCRIPTION
GERMAN TRANSCRIPTION

Figure 1: A language-identification (LID) system as a front end to a set of real-time speech recognizers. The LID system outputs its three best guesses of the language of the spoken message (in this case, German, Dutch, and English). Speech-recognizers are loaded with models for these three languages and make the final LID decision (in this case, Dutch) after decoding the speech utterance.

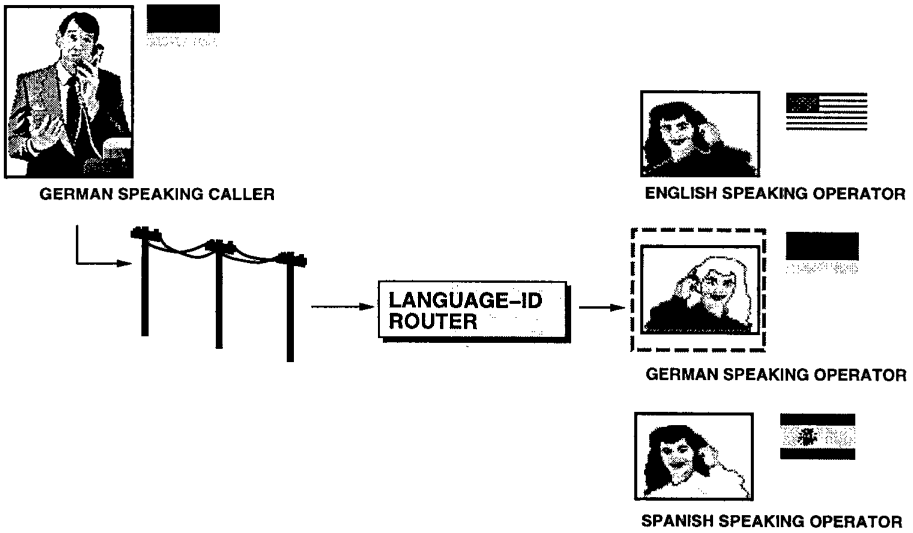


Figure 2: A language-identification (LID) system as a front end to a multi-lingual group of directory-assistance or emergency operators. The LID system routes an incoming call to a switchboard operator fluent in the corresponding language.

produced when a speaker is thinking of speaking a phoneme. The phones that comprise the world celebrate might be [s eh l ax bcl b r ey q]. As documented by linguists, phone and phoneme sets differ from one language to another, even though many languages share a common subset of phones/phonemes. Phone/phoneme frequencies of occurrence may also differ, i.e., a phone may occur in two languages, but it may be more frequent in one language than the other. Phonotactics, i.e., the rules governing the sequences of allowable phones/phonemes, can also be different.

- Morphology. The word roots and lexicons are usually different from language to language. Each language has its own vocabulary, and its own manner of forming words.

- Syntax. The sentence patterns are different among languages. Even when two languages share a word, e.g., the word "bin" in English and German, the sets of words that may precede and follow the word will be different.

- Prosody. Duration characteristics, pitch contours, and stress patterns are different from one language to another.

3. LANGUAGE IDENTIFICATION SYSTEMS

Research in automatic language identification from speech has a history extending back to the 1970s. A few representative LID systems are described below. The reader will find references to other LID systems in reviews by Muthusamy et al. [33] and Zissman [50].

Figure 3 shows the two phases of LID. During the "training" phase, the typical system is presented with examples of speech from a variety of languages. Each training speech utterance is converted into a stream of feature vectors. These feature vectors are computed from short windows of the speech waveform (e.g. 20 ms) during which the speech signal is assumed to be somewhat stationary. The feature vectors are recomputed regularly (e.g. every 10 ms) and contain spectral or cepstral information about the speech signal (the cepstrum is the inverse Fourier transform of the log magnitude spectrum; it is used in many speech processing applications). The training algorithm analyzes a sequence of such vectors and produces one or more models for each language. These models are intended to represent a set of language dependent, fundamental characteristics of the training speech to be used during the next phase of the LID process.

During the "recognition" phase of LID, feature vectors computed from a new utterance are compared to each of the language-dependent models. The likelihood that the new utterance was spoken in the same language as the speech used to train each model is computed and the maximum-likelihood model is found. The language of the speech that was used to train the model yielding maximum likelihood is hypothesized as the language of the utterance.

The key issue becomes that of modeling the languages. We will discuss a series of different features that have been extracted from speech, yielding increasing amounts of knowledge at the cost of rendering the language identifications system more and more complex. Some systems require only the digitized speech utterances and the corresponding true identities of the languages being spoken because the language models are based simply on the signal representation or on self generated token representation. More complicated LID systems use phonemes to model speech and may require either (1) a phonetic transcription (sequence of symbols representing the spoken sounds), or (2) an orthographic transcription (the text of the words spoken) along with a phonemic transcription dictionary (mapping of words to prototypical pronunciation) for each training utterance. Producing these transcriptions and dictionaries is an expensive, time consuming process that usually requires a skilled linguist fluent in the language of interest.

3.1. Spectral-Similarity Approaches

In the earliest automatic language ID systems, developers capitalized on the differences in spectral content among languages, exploiting the fact that speech spoken in different languages contains different phonemes and phones. To train these systems, a set of prototypical short-term spectra were computed and extracted from training speech utterances. During recognition, test speech spectra were computed and compared to the training prototypes. The language of the test speech was hypothesized as the language having training spectra that best matched the test spectra.

There were several variations on this spectral similarity theme. The training and testing spectra could be used directly as feature vectors, or they could be used instead to compute formant-based or cepstral features vectors. The training exemplars could be chosen either directly from the training speech or could be synthesized through the use of K-means clustering. The spectral-similarity could be calculated by the Euclidean, Mahalanobis, or some other distance metric. Examples of spectral similarity LID systems are those proposed and developed by Cimarusti [4], Foil [11], Goodman [13], and Sugiyama [45].

To compute the similarity between a test utterance and a training model, most of the early spectral-similarity systems calculated the distance between each test utterance vector and each training exemplar. The distance between each test vector and its closest exemplar was accumulated as an overall distance, and the language model having lowest overall distance was found. In a generalization of this vector quantization approach to LID, Riek [40], Nakagawa [37] and Zissman [49] applied Gaussian mixture classifiers to language identification. Here, each feature vector is assumed to be drawn randomly according to a probability density that is a weighted sum of multivariate Gaussian densities. During training, a Gaussian mixture model for the spectral or cepstral feature vectors is created for each language. During recognition, the likelihood of the test utterance feature vectors is computed given each of the training models. The language of the model having maximum likelihood is hypothesized. The Gaussian mixture approach is "soft" vector quantization, where more than one exemplar created during training impacts the scoring of each test vector.

Whereas the language identification systems described above perform primarily static classification, hidden Markov models (HMMs) [38], which have the ability to model sequential characteristics of speech production, have also been applied to LID. HMM-based language identification was first proposed by House and Neuburg [17]. Savic [41], Riek [40], Nakagawa [37], and Zissman [49] all applied HMMs to spectral and cepstral feature vectors. In these systems, HMM training was performed on unlabeled training speech. Riek and Zissman found that HMM systems trained in this unsupervised manner did not perform as well as some of the static classifiers that each had been testing, though Nakagawa eventually obtained bet-

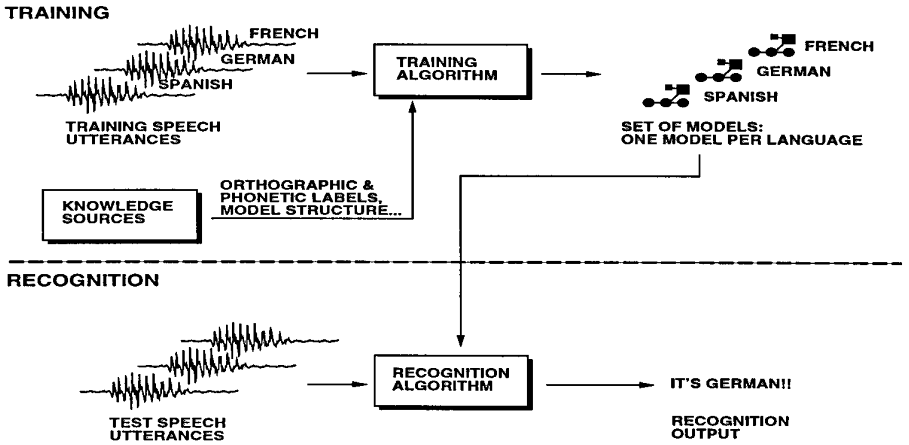


Figure 3: The two phases of language identification. During training, speech waveforms are analyzed and language-dependent models are produced. During recognition, a new speech utterance is processed and compared to the models produced during training. The language of the speech utterance is hypothesized.

ter performance for his HMM approach than his static approaches [36].

Li [26] has proposed the use of novel features for spectral-similarity LID. In his system, the syllable nuclei (i.e. vowels) for each speech utterance are located automatically. Next, feature vectors containing spectral information are computed for regions near the spectral nuclei. Each of these vectors is comprised of spectral sub-vectors computed on neighboring (but not necessarily adjacent) frames of speech data. Rather than collecting and modeling these vectors over all training speech, Li keeps separate collections of feature vectors for each training speaker. During testing, syllable nuclei of the test utterance are located and feature vector extraction is performed. Each speaker-dependent set of training features vectors is compared to the feature vectors of the test utterance, and the most similar speaker-dependent set of training vectors is found. The language of the speech spoken by the speaker of that set of training vectors is hypothesized as the language of the test utterance.

3.2. Prosody-based Approaches

Features that carry prosodic information have also been used as input to automatic language identification systems. This has been motivated, in part, by studies showing that humans can use prosodic features for identifying the language of speech utterances [35, 31]. For example, Itahashi has built systems that use features based on pitch estimates alone [18, 19]. He argues that pitch estimation is more robust in noisy environments than spectral parameters.

Hazen [14], however, showed that features derived from prosodic information provided little language discriminability when compared to a phonetic system. A system that used both prosodic and phonetic parameters performed about the same as a system using phonetic parameters alone.

Finally, Thyme-Gobbel et al. [47] have also looked at the utility of prosodic cues for language identification. Parameters were designed to capture pitch and amplitude contours on a syllable-by-syllable basis. They were normalized to be insensitive to overall amplitude, pitch and speaking rate. Results show that prosodic parameters can be useful for discriminating one language from another; however, the accuracy of any particular set of features is highly language-pair specific.

3.3. Phone-Recognition Approaches

Given that different languages have different phone inventories, many researchers have built LID systems that hypothesize exactly which phones are being spoken as a function of time and determine the language based on the statistics of that phone sequence. For example, Lamel built two HMM-based phone recognizers: one in English and another in French [25]. These phone recognizers were then run over test data spoken either in English or French. Lamel et al. found that the likelihood scores emanating from language-dependent phone recognizers can be used to discriminate between English and French speech. Muthusamy et al. ran a similar system on English vs. Japanese spontaneous, telephone-speech [32].

The novelty of these phone-based systems was the incorporation of more knowledge into the LID system. Both Lamel et al. and Muthusamy et al. trained their systems with multi-language phonetically labeled corpora. Because the systems require phonetically-labeled training speech utterances in each language, as compared to the spectral-similarity systems which do not require such labels, it can be more difficult to incorporate new languages into the language recognition process. This problem will be addressed further in Section 3.4.

To make phone-recognition-based LID systems easier to train, one can use a single-language phone recognizer as a front end to a system that uses phonotactic scores to perform LID. Phonotactics are the language-dependent set of constraints specifying which phonemes are allowed to follow other phonemes. For example, the German word "spiel" which is pronounced /sh p iy l/ and might be spelled in English as "shpeel" begins with a consonant cluster /sh p/ that cannot occur in English (except if one word ends in /sh/ and the next begins with /p/, or in a compound word like "flashpoint"). This approach is reminiscent of the work of D'Amore [9, 21], Schmitt [42], and Damashek [8], who have used n-gram analysis of text documents to perform language and topic identification and clustering. By "tokenizing" the speech message, i.e. converting the input waveform to a sequence of phone symbols, the statistics of the resulting symbol sequences can be used to perform language identification. Hazen [15] and Zissman [51] each developed LID systems that use one, single-language front end phone recognizer. An important finding of these researchers was that language ID could be performed successfully even when the front end phone recognizer(s) was not trained on speech spoken in the languages to be recognized. For example, accurate Spanish vs. Japanese LID can be performed using only an English phone recognizer. Zissman [51] and Yan [48] have extended this work to systems containing multiple, single-language front ends, where there need not be a front end in each language to be identified. Figure 4 shows an example of these types of systems.

3.4. Using Multilingual Speech Units

Alternative approaches to training language dependent phoneme recognizers use multi-lingual speech units. These are derived by either a mixture of language dependent and language independent phones or by deriving tokens automatically from training data. Advantages of this approach include data sharing and discriminant training between phonemes across languages and easy bootstrapping to unseen languages [10].

Research has also focused on the problem of identifying and processing only those phones that carry the most language discriminating information [1, 52]. These language-dependent phones are called "mono-phonemes" or "key-phones" in the literature. Kwan [24] and Dalsgaard [7] use both language specific and language independent phones in their systems. The language- independent phones, sometimes called "poly-phones", can be trained on data from more than one language without loss of language ID accuracy. Berkling [2], and Köhler [22, 23] have also tested systems that use a single multi language front end phone recognizer, i.e., a recognizer containing a mixture of "poly-phones" and "monophones".

3.5. Word Level Approaches

Between phone-level systems described in the previous sections and the large-vocabulary speech recognition systems described in a subsequent section are "word-level" approaches to language ID. These systems use more sophisticated sequence modeling than the phonotactic models of the phone-level systems, but do net employ full speech-to-text systems.

Kadambe [20] proposed the use of lexical modeling for language identification. An incoming utterance is processed by parallel language-dependent phone recognizers. Hypothesized language-specific word occurences are identified from the resulting phone sequences. Each language dependent lexicon contains several thousand entries. This is a bottom-up approach to the language ID problem, where phones are recognized first, followed by words, and eventually language. Thomas [46] has shown that a language-dependent lexicon need not be available in advance; rather, it can be learned automatically from the training data. Ramesh [39], Matrouf [29], Lund [28, 27] and Braun [3] have all proposed similar systems.

3.6. Continuous Speech Recognition

By adding even more knowledge to the system, researchers hope to obtain even better LID performance. Mendoza [30], Schultz [43, 44] and Hieronymus [16] have shown that large-vocabulary continuous-speech recognition systems can be used for language ID. During training, one speech recognizer per language is created. During testing, each of these recognizers is run in parallel, and the one yielding output with highest likelihood is selected as the winning recognizer—the language used to train that recognizer is the hypothesized language of the utterance. Such systems hold the promise of high quality language identification, because they use higher-level knowledge (words and word sequences) rather than lower-level knowledge (phones and phone sequences) to make the LID decision. Furthermore, one obtains a transcription of the utterance as a byproduct of LID. On the other hand, they require many hours of labeled training data in each language to be recognized and are the most computationally complex of the algorithms proposed.

4. EVALUATIONS

From 1993-1996, the National Institute of Standards and Technology (NIST) of the U.S. Department of Commerce has sponsored formal evaluation of language ID systems. At first, these evaluations were conducted using the Oregon Graduate Institute Multi-Language Telephone Speech (OGI-TS) Corpus [34]. The OGI-TS corpus contains 90 speech messages in each of the following 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. Each message is spoken by a unique speaker and comprises responses to ten prompts. For NIST evaluations, the monologue speech evoked by the prompt "Speak about any topic of your choice" is used for both training and testing. No speaker speaks more than one message or more than one language, and each speaker's message was spoken over a unique long-distance telephone channel. Pho-

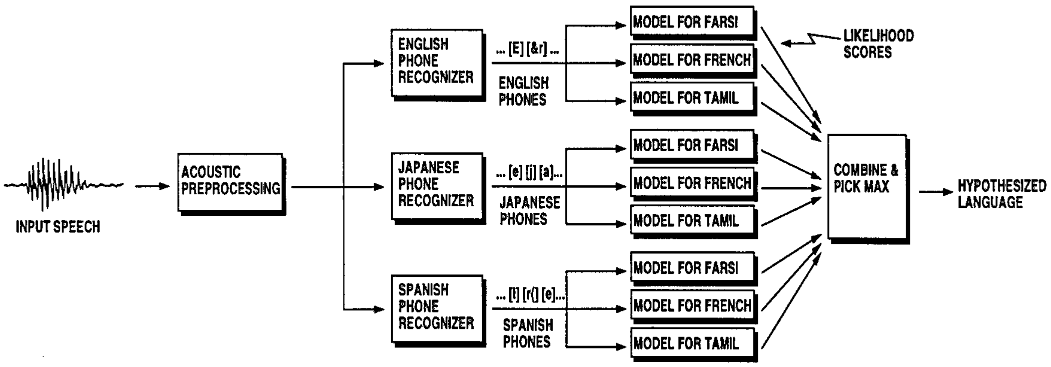


Figure 4: A LID system that uses several phone recognizers in parallel.

netically transcribed training data is available for six of the OGI languages (English, German, Hindi, Japanese, Mandarin and Spanish).

Performance of the best systems from the 1993, 1994 and 1995 NIST evaluations is shown in Figure 5. This performance represents each system's first pass over the evaluation data, which means that no system-tuning to the evaluation data was possible. For utterances having duration of either 45 s or 10 s, the best systems can discriminate between two languages with 4% and 2% error, respectively. This error rate is the average computed over all language pairs with English, e.g., English vs. Farsi, English vs. French, etc. When tested on nine-language forced-choice classification, error rates of 12% and 23% have been obtained on 45-s and 10-s utterances, respectively. The syllabic-feature system developed by Li and the systems with multiple phone recognizers followed by phonotactic language modeling developed by Zissman and Yan have exhibited the best performance over the years. Error rate has decreased over time, which indicates that research has improved system performance.

Starting in 1996, the NIST evaluations have employed the CALLFRIEND corpus of the Linguistic Data Consortium. CALLFRIEND comprises two-speaker, unprompted, conversational speech messages between friends. 100 North-American long distance telephone conversations were recorded in each of twelve languages (the same 11 languages as OGI-TS plus Arabic). No speaker occurs in more than one conversation. In the 1996 evaluation, the multiple phone recognizer followed by language modeling systems of Yan and Zissman performed best. The error rates on 30 s and 10 s utterances were 5% and 13% for pairwise classification. These same systems obtained 23% and 46% error rates for twelve-language classification. The higher error rates on CALLFRIEND are due to the informal conversational style of CALLFRIEND vs. the more formal monologue style of OGI-TS.

The CSR-based LID systems have not been fully evaluated at NIST evaluations, because orthographically and phonetically labeled speech corpora have not been available in each of the requisite languages. As such corpora become available in more languages, implementation and evaluation of CSR-based LID systems will become more feasible. Whether the performance they will afford will be worth their computational complexity remains to be seen.

5. CONCLUSIONS

Since the 1970s, language identification systems have become more accurate and more complex. Current systems can perform two-alternative forced-choice identification on extemporaneous monologue almost perfectly, and these same systems can perform 10-way identification with roughly 10% error. Though error rates on conversational speech are somewhat higher, there is every reason to believe that continued research coupled with competitive evaluations will result in improved system performance.

The improved performance of newer LID systems is due to their use of higher levels of linguistic information. Systems which try to model phones, phone frequencies, and phonotactics naturally perform better than those that model only lower-level acoustic information. Presumably, systems that model words and grammars will be shown to have even better accuracy.

Improved performance, however, comes at a cost. The higher levels of linguistic information must be programmed or trained into the newer LID systems. Whereas older systems required only digitized speech samples in each language to be recognized, more modern systems tend to require either a phonetic or orthographic transcription of at least some of the training utterances. State-of-the-art large-vocabulary CSR systems are often trained on hundreds of hours of transcribed speech. In recognition mode, these systems tend to run tens or even hundreds of times slower than real-time. Thus, the potential user of LID must balance the need for accuracy against the need for speedy deployment and low-cost (and possibly real-time) implementation.

6. REFERENCES

[1] K. M. Berkling, T. Arai, E. Barnard, and R.A.Cole. Analysis of phoneme-based features for language identification. In *International Conference on*

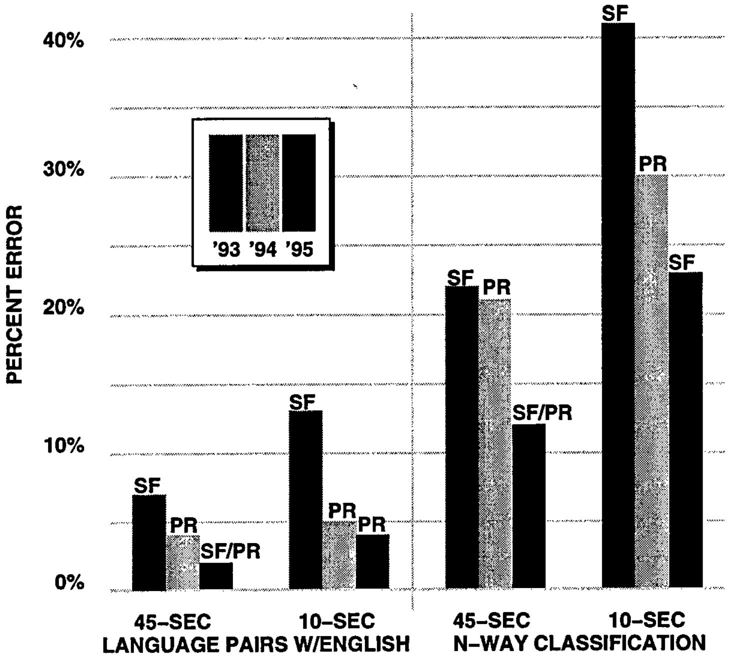


Figure 5: Error rates of the best LID systems at three NIST evaluations. Performance is shown on the left for average two-alternative, forced-choice classification of the various OGI-TS languages with English. "N-way" classification refers to 10-alternative, forced-choice performance in 1993, 11-alternative, forced-choice performance in 1994, and 9-alternative, forced-choice performance in 1995. "SF" indicates syllabic feature system. "PR" indicates phone recognition followed by language modeling system.

Acoustics, Speech, and Signal Processing, volume 1, pages 289 – 292, April 1994.

[2] K. M. Berkling and E. Barnard. Theoretical error prediction for a language identification system using optimal phoneme clustering. In *Eurospeech*, volume 1, pages 351–354, September 1995.

[3] J. Braun and H. Levkowitz. Automatic language identification with perceptually guided training and recurrent neural networks. In *International Conference on Spoken Language Processing*, volume 7, pages 3201–3205, October 1998.

[4] D. Cimarusti and R. B. Ives. Development of an automatic identification system of spoken languages: phase I. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1661–1663, May 1982.

[5] B. Comrie. *The World's Major Languages.* Oxford University Press, New York, 1990.

[6] D. Crystal. *The Cambridge Encyclopedia of Language.* Cambridge University Press, Cambridge, UK, 1987.

[7] P. Dalsgaard and O. Andersen. Identification of mono- and poly-phonemes using acoustic-phonetic features derived by a self-organizing neural network. In *International Conference on Spoken Language Processing*, pages 547–550, October 1992.

[8] M. Damashek. Gauging similarity with n-grams: language-independent categorization of text. *Science*, 267(5199):843–848, February 1995.

[9] R. J. D'Amore and C. P. Mah. One-time complete indexing of text: theory and practice. In *Proceedings of the Eighth Intl. ACM Conf. on Res. and Dev. in Information Retrieval*, pages 155–164, 1985.

[10] B. Wheatly et al. An evaluation of cross-language adaptation for rapid hmm development in a new langauge. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 237–240, April 1994.

[11] J. T. Foil. Language identification using noisy speech. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 861–864, April 1986.

[12] V. Fromkin and R. Rodman. *An Introduction to Language.* Harcourt Brace Jovanovich, Inc., Orlando, FL, 1993.

[13] F. J. Goodman, A. F. Martin, and R. E. Wohlford. Improved automatic language identification in noisy

speech. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 528–531, May 1989.

[14] T. Hazen. *Automatic Language Identification Using a Segment-based Approach*. PhD thesis, MIT, August 1993.

[15] T. J. Hazen and V. W. Zue. Automatic language identification using a segment-based approach. In *Eurospeech*, volume 2, pages 1303–1306, September 1993.

[16] J. L. Hieronymus and S. Kadambe. Robust spoken language identification using large vocabularly speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1111–1114, April 1997.

[17] A. S. House and E. P. Neuburg. Toward automatic identification of the language of an utterance. I. preliminary methodological considerations. *J. Acoust. Soc. Amer.*, 62(3):708–713, September 1977.

[18] S. Itahashi and L. Du. Language identification based on speech fundamental frequency. In *Eurospeech*, volume 2, pages 1359–1362, September 1995.

[19] S. Itahashi, J. Zhou, and K. Tanaka. Spoken language discrimination usin speech fundamental frequency. In *International Conference on Spoken Language Processing*, volume 4, pages 1899–1902, September 1994.

[20] S. Kadambe and J. Hieronymus. Language identification with phonological and lexical models. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3507–3511, May 1995.

[21] R. E. Kimbrell. Searching for text? Send an n-gram! *Byte*, 13(5):297–312, May 1988.

[22] J. Koehler. In-service adaptation of multilinual hidden-markov-models. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1451–1454, April 1997.

[23] J. Koehler. Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 417–420, April 1998.

[24] H.K. Kwan and K. Hirose. Use of recurrent network for unknown language rejection in language identification systems. In *Eurospeech*, volume 1, pages 63–67, September 1997.

[25] L. F. Lamel and J.-L. Gauvain. Cross-lingual experiments with phone recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 507–510, April 1993.

[26] K.-P. Li. Automatic language identification using syllabic spectral features. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 297–300, April 1994.

[27] M. A. Lund, K. Ma, and H. Gish. Statistical language identification based on untranscribed training data. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 793–796, May 1996.

[28] M.A. Lund and H. Gish. Two novel language model estimation techniques for statistical language identification. In *Eurospeech*, volume 2, pages 1363–1366, September 1995.

[29] D. Matrouf, M. Adda-Decker, L.F. Lamel, and J.L. Gauvain. Language identification incorporating lexical information. In *International Conference on Spoken Language Processing*, volume 2, pages 181–185, October 1998.

[30] S. Mendoza et al. Automatic language identification using large vocabulary continuous speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 785–788, May 1996.

[31] K. Mori, N. Toba, T. Harada, T. Arai, M. Komatsu, M. Aoyagi, and Y. Murahara. Human language identification with reduced spectral information. In *Eurospeech*, September 1999.

[32] Y. Muthusamy et al. A comparison of approaches to automatic language identification using telephone speech. In *Eurospeech*, volume 2, pages 1307–1310, September 1993.

[33] Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 11(4):33–41, October 1994.

[34] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *International Conference on Spoken Language Processing*, volume 2, pages 895–898, October 1992.

[35] Y. K. Muthusamy, N. Jain, and R. A. Cole. Perceptual benchmarks for automatic language identification. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 333–336, April 1994.

[36] S. Nakagawa, T. Seino, and Y. Ueda. Spoken language identification by ergodic HMMs and its state sequences. *Electronics and Communications in Japan, Part 3*, 77(6):70–79, February 1994.

[37] S. Nakagawa, Y. Ueda, and T. Seino. Speaker-independent, text-independent language identification by HMM. In *International Conference on Spo-*

ken Language Processing, volume 2, pages 1011–1014, October 1992.

[38] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[39] P. Ramesh and E. Roe. Language identification with embedded word models. In *International Conference on Spoken Language Processing*, volume 4, pages 1887–1890, September 1994.

[40] L. Riek, W. Mistretta, and D. Morgan. Experiments in language identification. Technical Report SPCOT-91-002, Lockheed Sanders, Inc., Nashua, NH, December 1991.

[41] M. Savic, E. Acosta, and S. K. Gupta. An automatic lanuguage identification system. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 817–820, May 1991.

[42] J. C. Schmitt. Trigram-based method of language identification. U. S. Patent 5,062,143, October 1991.

[43] T. Schultz, I. Rogina, and A. Waibel. LVCSR-based language identification. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 781–784, May 1996.

[44] T. Schultz and A. Waibel. Language independent and language adaptive large vocabulary speech recognition. In *International Conference on Spoken Language Processing*, volume 5, pages 1819–1823, October 1998.

[45] M. Sugiyama. Automatic language recognition using acoustic features. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 813–816, May 1991.

[46] H.L. Thomas, E. S. Parris, and J. H. Wright. Recurrent substrings and datafusion for language recognition. In *International Conference on Spoken Language Processing*, volume 2, pages 169–173, October 1998.

[47] A.E. Thyme-Gobbel and S.E. Hutchins. On using prosodic cues in automatic language identification. In *International Conference on Spoken Language Processing*, volume 3, pages 1768–1772, October 1996.

[48] Y. Yan and E. Barnard. An approach to automatic language identification based on language-dependent phone recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3511–3514, May 1995.

[49] M. A. Zissman. Automatic language identification using Gaussian mixture and hidden Markov models. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 399–402, April 1993.

[50] M. A. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Trans. Speech and Audio Proc.*, SAP-4(1):31–44, January 1996.

[51] M. A. Zissman and E. Singer. Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 305–308, April 1994.

[52] M.A. Zissman and E. Singer. Language identification using phoneme recognition and phonotactic language modeling. In *International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3503–3506, May 1995.

MULTILINGUAL TEXT-INDEPENDENT SPEAKER IDENTIFICATION

Geoffrey Durou

Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez
B-7000 Mons, Belgium
Email: durou@tcts.fpms.ac.be

ABSTRACT

In this paper, we investigate two facets of speaker recognition : cross-language speaker identification and same-language non-native text-independent spea-ker identification. In this context, experiments have been conducted, using standard multi-gaussian mo-deling, on the brand new multi-language TNO corpus. Our results indicate how speaker identification performance might be affected when speakers do not use the same language during the training and testing, or when the population is composed of non-native speakers.

1. INTRODUCTION AND MOTIVATION

Speaker recognition systems working in text independent (TI) mode have been characterized by their flexibility but also by their insecure aspect. Indeed, the non-imposing of words or sentences can lead to the breaking of the system if the voice of an authorized person is pre-recorded.

However, text-independent speaker identification systems are involved in many applications. That is the reason why many efforts have been developed in order to improve text-independent speaker recognition methods. For the last decade, the technology in this field has achieved significant progress. Now, these techniques can be used in real conditions, for that the application field be well defined.

Nowadays, more and more users of such systems are polyglot. So, if we do not have a priori know-ledge of the mother tongue of the talker - or at least the tongue he used during the training - and if we can not apply any language identification system, then it is possible to perform speaker identification in a language different from the one used during training. Let us note that no restriction about the tongue would still increase the flexibility of the system. However, the system may still impose one specific tongue. Since, it should be open to all users, we can easily imagine that any given language might differ from the native language of some of the users.

In order to start a descriptive study on (a) the cross-language and (b) the same non-native language effects on speaker recognition performance, we carried out some text-independent speaker identification experiments on a subset of 57 speakers extracted from the TNO multi-language database. Our system is based on the standard GMM technique, which has already been successfully used by the past for TI speaker recognition [3] [2] [4].

In section 2 we present in detail the TNO corpus and our identification system. The speaker identification experiments are described in section 3, which is subdivided into three items : (a) native speaker identification, acting as reference experiment; (b) cross-language speaker identification; (c) non-native same-language speaker identification. Results are then discussed and, in particular, cross-language spea-ker identification results are compared to performance recently obtained on the POLYCOST telephone speech corpus [5] [1].

2. EXPERIMENTAL SETUP

2.1. Database

Speech material for our experiments was taken from the new Dutch TNO corpus. This database consists in 82 Dutch speakers. All of them were prompted to pronounce 10 sentences in four different languages : Dutch, English, French, and German. All the sentences were read from a computer screen in a anechoic silent recording room. Given one language, the first five sentences are common for all speakers, while the others differ from one speaker to another.

We decided to accomplish the identification tests over all the speakers for whom speech data in the four tongues are available. So we conducted our experiments on a subset of 57 speakers (68 % males and 32 % females).

The first 5 utterances (per language identical for all speakers) were used for the training, while the other 5 sentences (per language and per speaker unique) were reserved to the identification tests.

In our experiments, we have systematically considered four different training durations (10 s, 15 s, 20 s, and 25 s) and five different testing durations (5 s, 10 s, 15 s, 20 s, and 25 s).

2.2. Feature Extraction

Speech recordings were sampled at 16 kHz. Analysis windows consisted of 512 samples taken every 16 ms. After pre-emphasis (factor 0.95) and application of a Hamming window, 10 autocorrelation LPC coefficient were computed and transformed into 12 cepstral coefficients. Finally, training and testing features consist only of 12 cepstral coefficients : neither the energy, nor dynamic information (delta coefficients), nor the pitch were used. No cepstral mean subtraction was applied.

2.3. Speaker Model

Our speaker identification system is based on the statistical modeling by Gaussian mixtures [3] [2] [4]. Each mixture is composed of 12 Gaussian distributions, with diagonal covariances matrices.

3. EXPERIMENTS

3.1. Native speaker Identification

First of all, let us carry out a preliminary experiment, considering both training and test phases in the mother tongue of the speakers. This might be seen, in the context of this paper, as the reference experiment.

Let us remind once again that for these experiments and all the experiments that will follow, we shall systematically choose the five sentences per language identical for the training, and the other five per language and per speaker unique for the identification tests.

The identification error rates for various training and testing durations are given hereafter in Figure 1.

We can notice at this point that the closed set speaker identification rate reaches 100 % for a 20 second testing duration and more, whatever the training duration considered.

3.2. Cross-language speaker identification

It would now be interesting to measure the impact of language on our speaker recognition system.

For that purpose, we conduct an experiment characterized by the use of different languages during the training and the test : models are trained on native speech (i.e. Dutch), while identification tests are made successsively on non-native speech (successively English, French, and German).

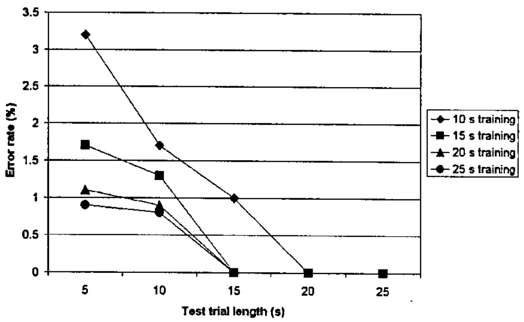


Figure 1: *Identification error rates over 57 native speakers of Dutch as a function of test trial length for various training conditions*

Results for different training and testing durations are reported in Figure 2, Figure 3 and Figure 4 below.

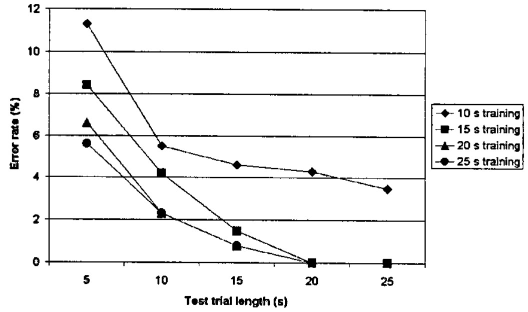


Figure 2: *Cross-language speaker identification error rates (Dutch / English) over 57 Dutch speakers as a function of test trial length for various training conditions.*

For values of training and testing durations large enough, we are still able, in the case Dutch/English, to reach the maximal performance.

On the contrary, we are unable to reach a 100 % identification rate in the case Dutch/French, given our proposed training and testing conditions.

When German is used for the test, error rates seem to converge to about 2 %.

Similar experiments have been recently conducted on a telephone speech database [1]. In this context, cross-language speaker identification tests on a set of 111 speakers showed that the performance degradation induced by

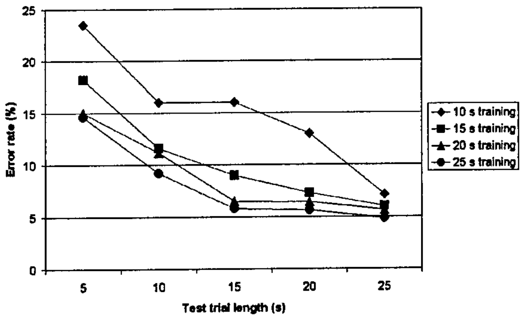


Figure 3: *Cross-language speaker identification error rates (Dutch / French) over 57 Dutch speakers as a function of test trial length for various training conditions.*

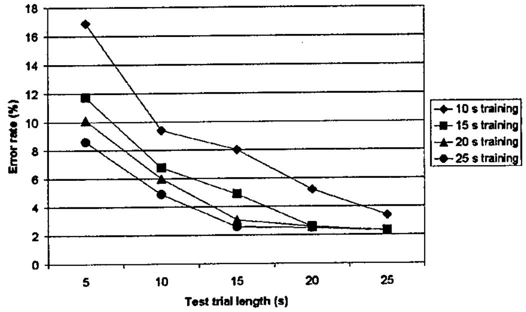


Figure 4: *Cross-language speaker identification error rates (Dutch / German) over 57 Dutch speakers as a function of test trial length for various training conditions.*

the use of a non-native tongue for the test did not exceed 1 % (relatively to the use of the native tongue for the test) in the case of a speaker identification system based on a vector quantization technique. We justified this very restricted difference by the fact that spectral characteristics of the speaker speech is not importantly modified as he speaks a second language. This corroborated another study which has shown that people who learn a second language at an advanced age (> 10 years old), instead of learning new phonemes, substitute phonemes from their native language and impose the rythm of this native language as they speak a non-native language [8]. Let us also mention that this conclusion was consolidated by an experiment described in [6] and which showed that the spectrum difference, measured by Kullback's divergence, on English and Japanese words pronounced by bilingual speakers was very small.

Here, in the case of maximal training and testing durations, we observe that the degradation easily exceeds 1 % in the cases Dutch-French (4.8 %) and Dutch-German (2.3 %) even though the population size is more restricted. However, we must be aware that, first, the maximal training duration is here of 25 seconds, whereas each training session lasted about 90 seconds in the previous work. Secondly, our identification system is now based on statistical modeling by Gaussian mixtures. These two points make it difficult to compare in the absolute results from these experiments.

3.3. Non-native speaker identification

Let us finally consider a last set of experiments conducted on non-native talkers. We conducted three sets of experiments characterized by the use of same non-native language during the training and the test : models were trained and identification tests were made on non-native speech (successively English, French, and German).

Once again, we report separately results on English, French, and German speech in Figure 5, Figure 6, and Figure 7, for different training and testing durations.

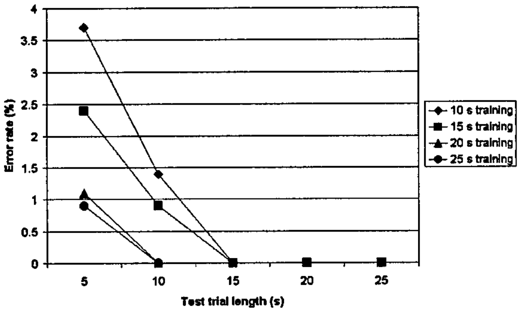


Figure 5: *Identification error rate over 57 non-native speakers of English as a function of test trial length for various training conditions.*

When English is chosen as non-native language, we see that there is no big difference between these plots and the reference plots. Surprisingly enough, the system performs sometimes better when this non-native language is employed.

We may reiterate the same observation if German is used. However, our system performs slightly worse if French is employed.

Globally, as expected, we observe through these experiments that even if non-native speakers use the phonetic and prosodic patterns of their first language, the identification scores are not really affected.

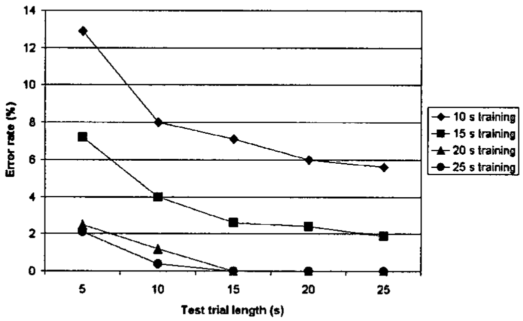


Figure 6: *Identification error rate over 57 non-native speakers of French as a function of test trial length for various training conditions.*

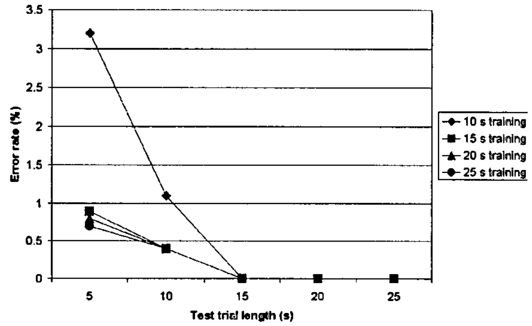


Figure 7: *Identification error rate over 57 non-native speakers of German as a function of test trial length for various training conditions.*

Major aspects that can make non-native speech deviate from native speech are notably fluency, word stress, and intonation [7]. Although these factors might be responsible of a score degradation in the cross-language case, we can easily understand that they have a much more restricted effect on these last experiments. In particular, if a non-native talker tends to speak more slowly during the training, he will also tend to speak roughly the same way for the tests, because the language is the same. This point should explain partly why the identification scores are not so affected.

4. CONCLUSION

The purpose of this paper was to describe and carry out multi-lingual speaker identification experiments on the TNO database made of native speakers of Dutch, and to comment on the results. Various training and testing durations were considered.

We first carried out a preliminary set of experiments (what we considered as being the baseline experiments) where both training of the speakers models and the identification tests were made on their mother tongue (i.e. Dutch). Then, regarding to our baseline results, we have measured the evolution of our speaker identification system performance when (a) different languages are used during the training and the tests; (b) a same non-native language is used both for the speakers models training and the identification tests. Three non-native languages were tested : English, French, and German.

We also pointed out and partly justified the discordance between the conclusions about the effect on the language if the performance degradation is measured on the microphone TNO corpus or on the telephone POLYCOST database.

5. REFERENCES

[1] G. Durou, F. Jauquet, "Cross-Language Text-Independent Speaker Identification", Proc. European Conference on Signal Processing (EUSIPCO'98), vol 3, pp 1481-1484, September 1998, Rhodes, Greece.

[2] D. A. Reynolds, "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification", PhD Thesis, Georgia Institute of Technology, 1992.

[3] G. McLachlan and K. Basford, "Mixture Models : Inference and Applications to Clustering", Marcel Dekker, 1998.

[4] D. Titterington, A. Smith, and U. Markov, "Statistical Analysis of Finite Mixture Distributions", John Wiley and sons, 1985.

[5] The European COST 250 action entitled "Speaker Recognition in Telephony", Information can be found on the web page : http://circhp.epfl.ch/polycost/

[6] M. Abe and K. Shikano, "Statistical analysis of bilingual speakers's speech for cross-language voice conversation", J. Acoust. Soc. Amer., Vol 90, pp 76-82, July 1991.

[7] C. Cucchiarini, H. Strik, and L. Boves, "Automatic evaluation of Dutch pronunciation by using speech recognition technology", Proc IEEE ASRU, Santa Barbara, Dec 1997.

[8] L. Neumeyer, H. Franco, M. Weintraub, and P. Price, "Automatic text-independent pronunciation scoring of foreign language student speech", Proc ICSLP'96, Philadelphia, pp 1457-1460, 1996.

VOWEL SYSTEM MODELING: A COMPLEMENT TO PHONETIC MODELING IN LANGUAGE IDENTIFICATION

François Pellegrino[1], *Jérôme Farinas*[2], *Régine André-Obrecht*[2]

[1]DDL - ISH
14 avenue Berthelot, 69 363 Lyon Cedex 07,
France
Francois.Pellegrino@univ-lyon2.fr

[2]IRIT University Paul Sabatier
118 route de Narbonne, 31 062 Toulouse,
France
(jfarinas,obrecht)@irit.fr

ABSTRACT

Most systems of Automatic Language Identification are based on phonotactic approaches. However, it is more and more evident that taking other features (phonetic, phonological, prosodic, etc.) into account will improve performances. This paper presents an unsupervised phonetic approach that aims to consider phonological cues related to the structure of vocalic and consonantal systems.

In this approach, unsupervised vowel/non vowel detection is used to model separately vocalic and consonantal systems. These Gaussian Mixture Models are initialized with a data-driven variant of the LBG algorithm: the LBG-Rissanen algorithm.

With 5 languages from the OGI MLTS corpus and in a closed set identification task, the system reaches 85 % of correct identification using 45-second duration utterances for male speakers. Using the vowel system modeling as a complement to an unsupervised phonetic modeling increases this performance up to 91 % while still requiring no labeled data.

1. INTRODUCTION

Until recently, Automatic Language Identification (ALI) was a marginal domain of automatic speech processing. The times are changing and today, it raises as one of the main challenges as far as Human-Computer Interfaces (HCI) are concerned. The need for multilingual capacities grows with the joined development of world communication and multi-ethnic societies as the European Economic Community. The language obstacle will remain until either multilingual large vocabulary continuous speech recognition or ALI systems reach excellent performance and reliability. Besides, video and audio contain-based indexing requires the extraction of extra linguistic information (music/speech segmentation, speaker and language identification).

Presently, the most efficient ALI systems are based on phonotactic discrimination via specific statistical language modeling [1,2,3,4]. In most of them, phonetic recognition is merely considered as a front-end: it consists in a projection of the continuous acoustic space into one or several discrete sets corresponding more or less to phonetic units. Though this approach achieves the best results, it seems that increasing performances necessitates to consider additional features (especially phonetic ones).

Obviously, if these features have been neglected for a while, it is because they are not so easy to exploit in ALI. Efficient phonetic modeling, based mainly on Hidden Markov models (HMM) used to require a consequent amount of hand-labeled data for training. Unfortunately, this kind of data is expensive to acquire and it is available only for a few languages (6 in the Multi Language Telephone Speech database from OGI [5]). Consequently, phonetic based systems were limited to these 6 languages. Fortunately, HMM reach today better performances and enhanced capacity of adaptation while requiring less and less hand-labeled data: phonetic modeling becomes a competitive approach and reaches good results [6].

Exploiting both phonetic and phonotactic cues is a very efficient approach, but we think that it may be significantly improved by taking phonology in consideration, especially for languages where no labeled data are accessible. For such languages, we propose to emphasize the structure of their phonological systems. This approach consists in two steps:

- splitting the speech utterance in segments corresponding with natural sound categories (vowels, fricatives, etc.) and then
- modeling each category as a whole, in order to capture the salient phonological cues of the language.

Linguists are collecting language descriptions and developing language typologies for a while [7]. We think that taking advantage of phonological typologies is a promising approach both for ALI and for automatic language description.

This paper reports experiments that aim to assess the discriminative power of an unsupervised phonological approach.

Next section will describe briefly the two systems (a global segmental model or GSM and a Phonetic Differentiated Model or PDM) which are used in the experiments. Each model is then described in details (Sections 3 and 4). Experiments on the OGI MLTS database are reported in Section 5. We discuss the performance and the perspective of such approaches in the conclusion paragraph.

2. DESCRIPTION OF THE SYSTEMS

Two systems have been implemented for these experiments.

In the first one, all the utterances of a given language are segmented, gathered and modeled by a single Gaussian Mixture Model (GMM) evaluated in a cepstral space. This Global Segmental Model (GSM) is partially similar to the simplest model proposed by M. Zissman in [4] and is used as a reference system.

The second model is an extension of the first one, but it is designed to test the hypothesis that the structural information on the vowel system of each language can be modeled to identify it. A vowel detection algorithm is used to split the segments gathered for each language in 2 categories: vowel and non-vowel. For each language, one GMM is subsequently evaluated from each set: a Vowel System Model (VSM) and a Consonantal System Model (CSM) though non-vowel segments can not be exactly considered as consonants (vowel transitions may also be labeled as non-vowels).

The choice of the vowel/non-vowel distinction is based on both linguistic and acoustic considerations: from a linguistic point of view, vowel system typologies are available for a few years [8]. Additionally, the homogeneous structure of the vocalic acoustic space provides a good framework to investigate structure modeling.

Both systems take advantage from an a priori segmentation algorithm [9]. It provides variable length segments by detecting ruptures in the statistical structure of the speech signal. This way, a duration information is provided for each sound before any additional modeling.

3. GLOBAL SEGMENTAL MODEL

The idea of modeling all the sounds of a language in a single model is not new. It has been first proposed in the 80's and M. Zissman has implemented it in [4]. The goal is to model the phonetic space of each language rather than each phone. The advantage is that it does not require any knowledge on the allophones for each language. Unfortunately, it tends to be less discriminative than a phone modeling approach. However, taking the duration provided by the a priori segmentation into account may enhance the performances as it is used to in speech recognition [10].

3.1 Statistical framework

Let L = $\{L_1, L_2, \ldots, L_{NL}\}$ be the set of N_L languages to identify; the problem is to find the most likely language L^* in L, given that the effective language is really in this set (closed set experiments).

Let T be the number of segments in the spoken utterance and $O = \{o_1, o_2, \ldots o_T\}$ the sequence of observation vectors. Given O and using Bayes' theorem, the most likely language L^* according to the model is:

$$L^* = \underset{1\le i\le N_L}{\arg\max}\left[\Pr(L_i|O)\right] = \underset{1\le i\le N_L}{\arg\max}\left[\frac{\Pr(O|L_i)\Pr(L_i)}{\Pr(O)}\right]$$

$$L^* = \underset{1\le i\le N_L}{\arg\max}\left[\Pr(O|L_i)\Pr(L_i)\right] \quad (1)$$

Additionally, if *a priori* language probabilities are assumed to be identical, one gets the equation:

$$L^* = \underset{1\le i\le N_L}{\arg\max}\left[\Pr(L_i|O)\right] = \underset{1\le i\le N_L}{\arg\max}\left[\Pr(O|L_i)\right] \quad (2)$$

Under the standard assumptions, each segment is considered independent of the others, conditionally to the language model. Finally, L^* is given in the log-likelihood space by:

$$L^* = \underset{1\le i\le N_L}{\arg\max}\left[\sum_{k=1}^{T}\log\Pr\left(o_k|L_i\right)\right] \quad (3)$$

For each language $\boldsymbol{L_i}$, a GMM is trained with the set of speech segments. The EM algorithm is used to obtain the maximum likelihood parameters of each model [11]. This algorithm presupposes that the number of the mixture components, Q_i, and initial values for each Gaussian probability density functions are given; in our system, the LBG [12] and/or the LBG Rissanen algorithms [13] fix these parameters. During the recognition, the utterance likelihood is computed with the speech segments according to each language-specific model.

3.2 GSM Implementation

The training procedure consists in the following processing:

- An a priori segmentation provides steady and transient segments.
- A speech activity detector is applied to discard pauses.
- A cepstral analysis is performed on each segment.
- One GMM per language is estimated with the set of language dependent observations.

Note that, unlike most acoustic-phonetic decoders, the cepstral analysis is performed on variable length segments rather than on constant duration frames; the segment duration is added to the observation vector.

The same acoustic processing is applied during recognition, and the language is identified via a maximum likelihood computation of the utterance according to the language dependent models.

3.2.1 Segmentation and speech activity detection

The segmentation is provided by the "Forward-Backward Divergence" algorithm [9], which is based on a statistical study of the acoustic signal. Assuming that the speech signal is described by a string of quasi-stationary units, each one is characterized by an auto regressive Gaussian model; the method consists in performing an on line detection of changes in the auto regressive parameters. The use of this segmentation partially removes redundancy for long sounds, and a segment analysis is very useful and relevant to locate coarse features.

The segmentation is followed by a Speech Activity Detection in order to discard pauses. Each segment is labeled "silence" or "speech"; long silences (longer than 150 ms) are considered as non-speech and subsequently discarded.

3.2.2 Cepstral analysis

A set of 8 Mel-Frequency Cepstral Coefficients (MFCC) and 8 delta-MFCC characterize each segment. Cepstral analysis is performed using a 256-point Hamming window centered on the segment. This parameter vector may be extended with the duration of the underlying segment. A cepstral subtraction performs blind deconvolution (to remove the channel effect) and speaker normalization.

3.2.3 GMM Modeling

- *Initializing GMM with the LBG algorithm*

The LBG algorithm [12] elaborates a partition of the observation space by performing an iterated clustering of the learning data into codewords optimized according to the nearest neighbor rule. The splitting procedure may be stopped either when the data distortion variation drops under a given threshold or when a given number of codewords is reached. This last procedure has been used in our experiments.

- *Initializing GMM with the LBG Rissanen algorithm*

The LBG-Rissanen algorithm is similar to the LBG algorithm except for the iterated procedure termination. Before splitting, the Rissanen criterion *J(q)* [13, 14], function of the size *q* of the current codebook is computed from the expression:

$$J(q) = D_q(X) + 2p.q.\log(\log N) \qquad (4)$$

In this expression, $D_q(X)$ denotes the log-distortion of the training set *X* according to the current codebook, *p* the parameter space dimension and *N* the cardinal of *X*. Minimizing *J(q)* results in the optimal codebook size according to the Rissanen information criterion. We use this data-driven algorithm to determinate automatically the optimal number Q_i of Gaussian pdfs for each language.

3.2.4 Recognition processing

During the identification phase, the utterance is processed the same way, and its likelihood is computed according each language model using the speech segments. According to equation (3), the maximum likelihood rule is applied.

4. PHONETIC DIFFERENTIATED MODEL

In the PDM approach, language independent vowel detection is performed prior to the cepstral analysis. The detection locates segments that match vowel structure according to an unsupervised language-independent algorithm [15]. For each language $\boldsymbol{L_i}$, a Vowel System GMM, $\boldsymbol{VS_i}$, (respectively a Consonantal System GMM, $\boldsymbol{CS_i}$) is trained with the set of detected vowel segments (resp. non-vowel segments).

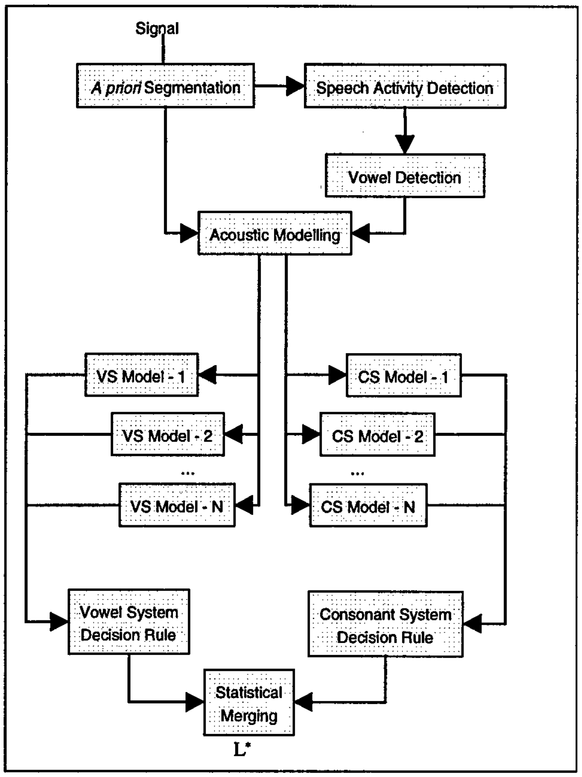


Figure 1 - Block diagram of the Phonetic Differentiated Model system. The upper part represents the acoustic preprocessing and the lower part the language dependent Vowel and Consonant-System Modeling.

4.1 Statistical framework

Let *T* be the number of segments given by the segmentation in the spoken utterance and $O = \{o_1, o_2, \ldots o_T\}$ be a sequence of observation vectors. Each vector o_k consists of a cepstral vector y_k and a macro-class flag c_k, equal to 1 if the segment is detected as a vowel, and equal to 0 otherwise. In order to simplify the formula, we note $o_k=\{y_k, c_k\}$.

Since (c_k) is a deterministic process, the most likely language computed in the log-likelihood space is given by:

$$L^* = \underset{1 \le i \le N_L}{\operatorname{argmax}} \left\{ \left[\sum_{c_k=1} \log \Pr(y_k | VS_i) \right] + \left[\sum_{c_k=0} \log \Pr(y_k | CS_i) \right] \right\} \quad (5)$$

4.2 PDM Implementation

Vowel detection is based on a spectral analysis algorithm. It is language independent and no training procedure is required.

To train the VS and CS models, the procedure is the same as the one used for training the GSM. The EM algorithm is combined with an initialization, by the LBG algorithm or the LBG-Rissanen algorithm.

In recognition phase, the utterances are processed the same way. It provides two sets of observations (vowel and non-vowel segments). For each language, two likelihoods are computed, according to the VS and the CS models. The maximum likelihood rule is applied to the overall likelihood (computed according to equation 5).

5. EXPERIMENTS

5.1 Corpus description

The OGI MLTS corpus [5] has been used in our experiments. The study is currently limited to 5 languages (French, Japanese, Korean, Spanish and Vietnamese). The phonological differences of the vowel system between these languages have motivated the use of this subset [8]. Spanish and Japanese vowel systems are rather elementary (5 vowels) and quasi-identical. Korean and French systems are quite complex, and they make use of secondary articulations (long vs. short vowel opposition in Korean and nasalization in French). Vietnamese system is of average size.

The aim of this corpus is to estimate the discriminative power of vowel system modeling with either close phonological VS or different ones, when salient features are available (e.g. nasal vowels).

The data are divided into two corpora, namely the training and the development sets. Each corpus consists in several utterances (constrained and unconstrained). There are about 20 speakers per language in the development subset and 50 speakers per language in the learning one. There is no overlap between the speakers of each corpus. The identification tests are made with a subset of the development corpus, called '45s' set, since 45s is the mean duration of the utterances.

5.2 Global Segmental Model

Several acoustic analyses and the two procedures of initialization have been assessed with the GSM system. Preliminary experiments have shown that considering the segment duration always improves performances. With 5 languages, the correct identification rate raises 86 % using the classical LBG algorithm initialization with the codebook size constrained.

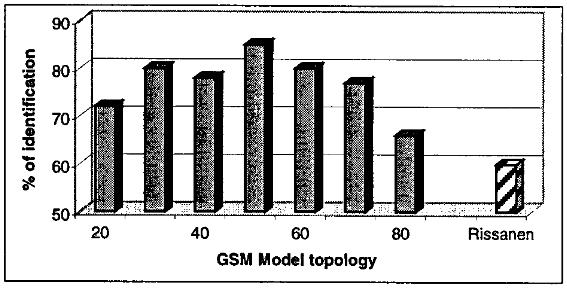


Figure 2 – Correct identification rate as a function of the GSM model topology. Dash bar corresponds with GSM initialized by LBG-Rissanen and plain bars with LBG algorithm (the a priori codebook size is displayed).

These results are obtained with 50 Gaussian laws for each language. The LBG-Rissanen algorithm is quite inefficient (see Figure 2). It does not handle correctly with the complexity of the global acoustic space and it is trapped, resulting in ineffective codebook sizes smaller than the expected ones (see Table 1).

5.3 Phonetic Differentiated Model

5.3.1 Vowel system modeling

To assess the VS models, a first sequence of experiments has been performed: the most likely language L^* is computed according to the VS models and non-vowel segments are discarded. When using the LBG algorithm, the best result is 67 % of correct identification (with 20 Gaussian components by VS model). Using the LBG-Rissanen algorithm to estimate the optimal size of each VS GMM is more efficient since the identification rate reaches 78 % (Figure 3). Remembering that only vowel segments are used (i.e. less than 10 seconds per utterance), this result shows that the VSM coupled with the LBG-Rissanen algorithm is able to correctly capture the structure of the vowel systems unlike what happened with GSM. Codebook sizes determined by LBG-Rissanen are significantly higher and the joined performances are much better for VSM than for GSM (see Table 1).

	French	Japanese	Korean	Spanish	Vietnamese
GSM	15	12	12	20	10
VSM	29	24	23	22	21
CSM	22	23	24	25	27

Table 1: Language-dependent model size given by LBG-Rissanen algorithm as a function of the parameter set (global, vocalic or consonantal).

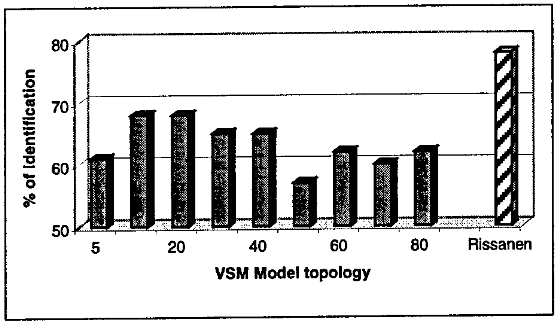


Figure 3 – Correct identification rate as a function of the **VSM** model topology. Dash bar corresponds with VSM initialized by LBG-Rissanen and plain bars with LBG algorithm (the a priori codebook size is displayed).

5.3.1 Consonant system modeling

The same kind of experiments has been performed to assess the CS models. Non-vowel segments are used (about 25 seconds per utterance). The best performance has resulted from the initialization of the GMM with the LBG algorithm: 30 Gaussian models reach 78 % of correct identification (Figure 3). The LBG-Rissanen algorithm has provided less discriminative models than those of constant size: consonant segments are acoustically more heterogeneous than vowel segments. Therefore, the consonant parameter space is much more complex than the vowel space and the LBG-Rissanen is unable to deal with it, similar to its behavior with the GSM.

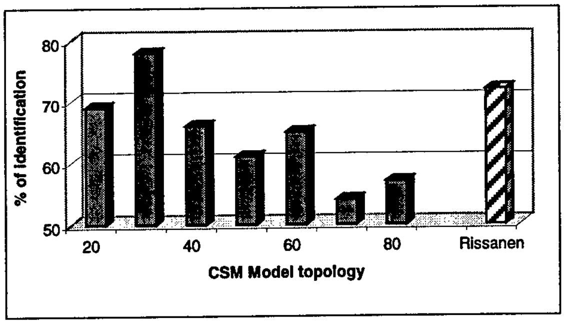


Figure 4 – Correct identification rate as a function of the **CSM** model topology. Dash bar corresponds with GSM initialized by LBG-Rissanen and plain bars with LBG algorithm (the a priori codebook size is displayed).

5.3.1 Phonetic Differentiated Modeling

The previous CS and VS models are combined to give the PDM approach (equation 5); The best system merges the VS model initialized by the LBG Rissanen algorithm and the CS model initialized by the classical LBG. 85 % of correct identification is reached.

5.4 GSM and PDM Comparison

As the previous experiments have shown, no significant differences, in term of identification rate, arises between the PDM and GSM approaches since they reach respectively 85% and 86% of correct identification (Table 2).

VSM	CSM	PDM	GSM
78	78	85	86

Table 2: Identification scores with all languages among 5 languages (45s male utterances).

In order to see if the information extracted from the signal by the two approaches is redundant or complementary, another sequence of experiments is performed to merge the different models. Scores provided by the considered models are combined and the maximum score is selected.

The best performance is reached when the GSM system and the VS model system are merged: identification rate among 5 languages raises from 86 % to 91 % (see Figure 5). The combination "CS model–GSM" does not improve the results: consonantal information seems to be redundant with GSM ones. When we merge the results of the GSM and the PDM, the results are intermediate: the CS modeling attenuates the gain of the VS modeling.

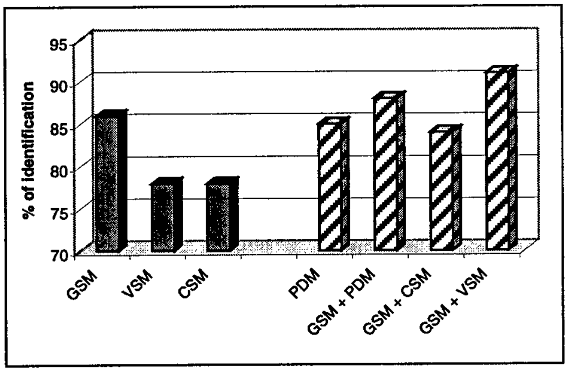


Figure 5 – percentage of correct identification according to the models. Dash bars correspond with systems resulting from merging.

The improvement of performance when using VSM as a complement to GSM is statistically significant. Additional experiments have been done to investigate if it is due to the redundant use of the vowel segments (resulting in a double weight with respect to consonantal segments) or if the VSM brings additional information. They confirm that the improvement is not an artifact of the weighting factor applied to vowel segments. Thus, the structure of the vowel system is a discriminative feature that is complementary to global phonetic modeling.

Additional experiments have been done with 3 languages, in order to compare with systems proposed in the literature. The figure 6 shows the results for the male part of the test corpus and for the global test set. The mean results are respectively 93.3 % and 86.4 %. This last result must be compared to the 84% obtained by O. Andersen [16] and 91% by S. Kadambe [17]. In these systems, Hidden Markov Models (HMM) and n-

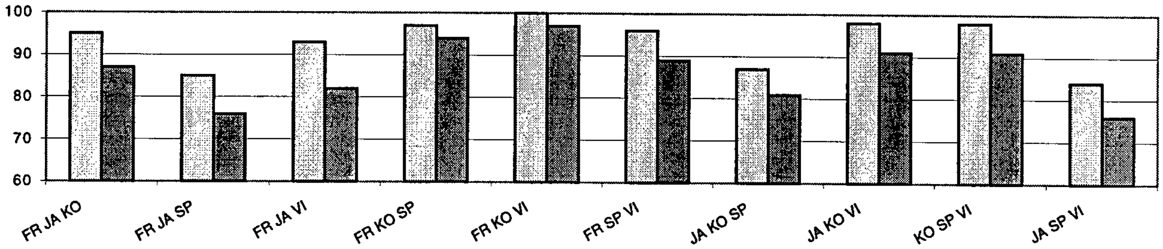


Figure 6: Identification rate for a 3 language identification task, and the '45s' test set. (in light, the test is limited to the male speaker set, while in dark, both male and female speakers are considered). Note that the models have been trained only with male speakers.

gram models have been used to model respectively the acoustic space and the phonotactic level.

6. CONCLUSION

This work proves that a significant part of the language characterization is embedded in its vowel system: vowel segments seem to be highly discriminative since the same level of performance is reached with vowel system modeling and consonantal system modeling though the consonantal duration is twice the vocalic duration in the utterances. Moreover, vowel system modeling using the LBG-Rissanen algorithm provides additional identification cues that are not exploited in the global segmental model (GSM). Thus, merging of the GSM and the VSM shows that extracting and modeling this information is possible and efficient.

The interest of the differentiated modeling approach is actual, and many advantages of the use of acoustic modeling in homogeneous spaces may be pointed out:

- Minimum Description Length algorithms (like LBG-Rissanen) are able to handle with the structure of the acoustic-phonetic system.
- A better discrimination is reached *inside* each model.
- The parameter space can be adapted to the characteristics of the acoustic class that is modeled

We will complete the notion of differentiated model, by introducing different model structures (GMM, HMM) and different acoustic parameters dependent of the phonetic classes (occlusive, fricative, et al). Then, to compare this approach to the classical ones, it will be necessary to complete our system with a phonotactic model, appropriate to our own acoustic projection.

7. REFERENCES

[1] T. J. Hazen, & V. W. Zue, (1997), Segment-based automatic language identification, *Journal of the Acoustical Society of America*, Vol. 101, No. 4, pp. 2323-2331, April.

[2] L.F. Lamel, J.L. Gauvain, (1994), Language Identification using Phone-Based Acoustic Likelihood, *Proc. of ICASSP '94*, Adelaide, pp. 293-296.

[3] Y. Yan, E. Barnard & R. A. Cole, (1996), Development of An Approach to Automatic Language Identification based on Phone Recognition, *Computer Speech and Language*, Vol. 10, n° 1, pp 37-54, (1996)

[4] M.A. Zissman, (1996), Comparison of four approaches to automatic language identification of telephone speech. *Proc. IEEE Trans. on SAP*, January 1996, vol. 4, n° 1.

[5] Y. K. Muthusamy, R. A. Cole & B. T. Oshika, (1992), The OGI Multilingual Telephone speech Corpus, *Proc. of ICSLP '92*, Banff, pp. 895-898

[6] D. Matrouf, M. Adda-Decker, J.-L. Gauvain & L. Lamel, (1999), Identification automatique de la langue par téléphone, actes de la *1ère Journée d'étude du GFCP sur l'identification automatique des langues*, Lyon.

[7] I. Maddieson, (1986), *Patterns of sounds*, 2nd Edition, Edited by Cambridge Univ. Press, USA

[8] N. Vallée, (1994), *Systèmes vocaliques : de la typologie aux prédictions*, Thèse de 3ème cycle, Univ. Stendhal, Grenoble

[9] R. André-Obrecht, (1988), A New Statistical Approach for Automatic Speech Segmentation. *IEEE Trans. on ASSP*, January 88, vol. 36, n° 1.

[10] R. André-Obrecht, B. Jacob, (1997), Direct Identification vs. Correlated Models to Process Acoustic and Articulatory Informations in Automatic Speech Recognition, *Proc. of ICASSP '97*, Munich, pp. 989-992.

[11] A.P. Dempster, N.M. Laird, D.B. Dubin, (1977), Maximum likelihood from incomplete data via the EM algorithm, *J. Royal statist. Soc. ServB.*,39.

[12] Y. Linde, A. Buzo, R.M. Gray, (1980), An algorithm for vector quantizer. *IEEE Trans on Com.*, January 80, vol 28.

[13] J. Rissanen, (1983), An universal prior for integers and estimation by minimum description length. *The Annals of statistics*, vol 11, n° 2.

[14] N. Parlangeau, F. Pellegrino and R. André-Obrecht (1999), Investigating Automatic Language Discrimination via Vowel System And Consonantal System Modeling, *Proc. of ICPhS '99*, San Francisco.

[15] F. Pellegrino, R André-Obrecht, (1997), From vocalic detection to automatic emergence of vowel systems, *Proc. ICASSP'97*, Munchen, April 1997.

[16] O. Andersen & P. Dalsgaard, Language-Identification Based on Cross-Language Acoustic Models and Optimised Information Combination, *Proc. of Eurospeech '97*, Rhodes, pp. 67-70, (1997)

[17] S. Kadambe, J.L. Hieronymous, (1994), Spontaneous speech language identification with a knowledge of linguistics, *Proc. of ICSLP'94*, Yokohama, pp. 1879-1882.

SCoPE, SYLLABLE CORE AND PERIPHERY EVALUATION: AUTOMATIC SYLLABIFICATION AND APPLICATION TO FOREIGN ACCENT IDENTIFICATION

Kay Berkling

M.I.T. Lincoln
Laboratory, 244 Wood Street,
Lexington, MA 02420-9185, USA
kay@sst.ll.mit.edu

Julie Vonwiller, Chris Cleirigh

University of Sydney, Dept. of Elect. Eng.
Sydney, Australia,
julie,cleirig@speech.su.oz.au

ABSTRACT

In this paper we apply a study of the structure of the English language towards an automatic syllabification algorithm. Elements of syllable structure are defined according to both their position in the syllable and to the position of the syllable within word structure. Elements of syllable structure that only occur at morpheme boundaries or that extend for the duration of morphemes are identified as peripheral elements; those that can occur anywhere with regard to word morphology are identified as core elements. All languages potentially make a distinction between core and peripheral elements of their syllable structure, however the specific forms these structures take will vary from language to language. In addition to problems posed by differences in phoneme inventories, we expect speakers with the greatest syllable structural differences between native and foreign language to have greatest difficulty with pronunciation in the foreign language. In this paper we will analyse two accents of Australian English: Arabic whose core/periphery structure is similar to English and Vietnamese, whose structure is maximally different to English.

1. INTRODUCTION

The goal of this paper is to exploit detailed knowledge of the English syllable structure model in order to add another dimension to phoneme-based feature analysis of foreign accented speech. This application to foreign accented speech in English derives from a more general study of the syllable structure of languages. The first part of this paper is therefore devoted to the application of this study to English, followed by an analysis of foreign accents in English as a function of syllable position. Properties of accented speech are expressed in terms of phoneme substitutions, deletions or insertions as a function of syllable position. A very simple example of the importance of position is provided by German phonology. Speakers tend to devoice obstruents (stops, fricatives and affricates) at ends of words but rarely in the middle. Position independent substitution probabilities would be inaccurate for both cases. By meaningfully discriminating position of the phoneme, we can potentially improve our feature set (of phoneme substitutions) for this type of phonological variation. In this paper we will analyse two accents of Australian English: (1) Arabic whose syllable structure is relatively similar to English. (2) Vietnamese, whose syllable structure is considerably different to that of English. Section 2 will describe an automatic syllabification algorithm of a pronunciation dictionary followed by a syllable structure analysis. Section 3 will analyse the differences in pronunciation as a function of syllable position for both foreign accents.

2. ENGLISH SYLLABLE STRUCTURE

Syllabification of pronunciation dictionaries is an important problem because syllable information is used for text to speech synthesis and can be an important feature in speech recognition. Most theoretical approaches to syllabification take the beginning or ending of words as their guide to the sorts of syllable structures that are allowable in a given language. In contrast, this paper takes morpheme-internal syllable structures as the basic template, and treats syllable structures specific to morpheme boundaries as exceptional, inasmuch as they carry boundary information. In order to understand the syllabification algorithm that is used in this work, we first present the model of syllable structure and the rationale that motivates it.

2.1. Syllable Constituents

A syllable usually consists of an obligatory vowel with optional surrounding consonants the exception being where a schwa-like vowel and following consonant are realised singly as a syllabic consonant. One familiar way of subdividing a syllable is into *Onset* and *Rhyme*. However, these categories alone do not indicate where the syllable

THIS WORK WAS SUPPORTED IN PART BY TWO CONSECUTIVE POST-DOC POSITIONS AT SYDNEY UNIVERSITY AND PROF. FURUI'S LABORATORY AT TOKYO INSTITUTE OF TECHNOLOGY, AND IN PART BY THE DEPARTMENT OF THE AIR FORCE. OPINIONS, INTERPRETATIONS, CONCLUSIONS, AND RECOMMENDATIONS ARE THOSE OF THE AUTHORS AND ARE NOT NECESSARILY ENDORSED BY THE UNITED STATES AIR FORCE.

is placed within the word. We propose another additional structure of the syllable as shown in Figure 1 which distinguishes between a *Core* and *Periphery*.

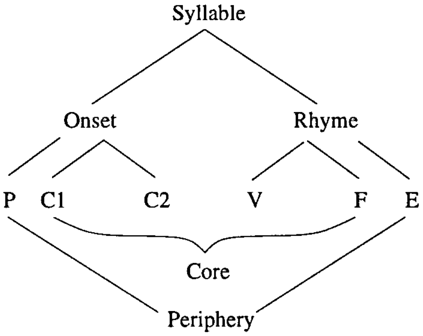


Figure 1: Constituents of a syllable as defined in this paper. (P, $C1$, $C2$, F, and E denote allowed sets of consonants. V denotes the set of vowels.)

In English, peripheral phonemes are those consonants that only occur as syllable constituents at morpheme boundaries. As such, the Periphery is a marker of morphological boundaries, and more often than not, this means word boundaries. We take the Periphery to be essentially a word-boundary phenomenon that can come to be incorporated within words historically through such processes as compounding. As an example, the word "flame" (*/fleim/*) can be broken down into the constituents as */flei/* (Core) and */m/* (Periphery), where the periphery demarcates the end of the (monomorphemic) word. Similarly, the word "lodgement" (*/lOdZm@nt/*) contains two syllables, */lOdZ/* and */m@nt/*; the first syllable has */lO/* (Core) and */dZ/* (Periphery), while the second has */m@n/* (Core) and */t/* (Periphery). Here the first Periphery */dZ/* marks the end of the first morpheme "lodge", and the second Periphery */t/* marks both the end of the second morpheme "-ment", and the end of the word "lodgement". By way of contrast, the word "freely" (*/fri:li:/*) contains two syllables, */fri:/* and */li:/*; the first syllable has */fri:/*(Core), while the second has */li:/*(Core). In this case then, although this word contains two morphemes, free and -ly, neither is demarcated by peripheral elements of syllable structure. While all languages potentially make a distinction between core and peripheral elements of their syllable structure, these structures will vary from language to language. Where English has demarcative consonants at syllable boundaries as Periphery, for tone-languages, such as Vietnamese, it is the "lexical" tone, which extends for the duration of the morpheme or word, that is analysed as the peripheral element of syllable structure. By analysing syllables in this way, we are able to identify not just differences in phoneme inventories across languages, but also differences in the ways that languages position their phonemes in syllables, and, importantly, differences in the ways that languages vary syllable structure according to the morphological location of a syllable. Comparing languages using such fine distinctions provides us with a powerful predictive tool for identifying elements of syllable structure that should prove most difficult for foreign speakers of English, and as such, a rich theoretical resource for the automated recognition of foreign accents of English.

2.2. Syllable Marking

In order to use the linguistic knowledge of syllable constituents as defined, we now want to devise an automatic method of marking syllables. Each pronunciation of a dictionary which is used by the system, will have to be split, first into syllables and then into its constituents. There are some basic rules for splitting a word into syllables. At the nucleus of any syllable is always the vowel (syllabic consonants are treated here as /@/+ consonant); long vowels and diphthongs count as a single phoneme, but occupy two syllable positions (V+F). Considering syllable structure in terms of the constituents Onset and Rhyme, the Rhyme begins with the vocalic nucleus, and anything before it in the same syllable is the Onset, a complex Onset being one containing more than one consonant. If there is only one consonant between two vowels, then that consonant is the Onset of the second syllable. If there are two consonants abutting of the same sonority, the syllable boundary falls between them, as in "threadbare." In general, if there are several consonants between vowels, then the consonant with the lowest sonority marks the start of the second syllable. The sonority hierarchy is given in Table 1 [3]. The principal exception to this is peripheral */s/*. For example, in the compound word "snakeskin" */sneikskIn/*, the word-internal proclitic */s/* that starts the second syllable falls between two consonants (*/k/*) of lower sonority Note that, on phonological criteria alone, it is not possible to determine whether peripheral */s/* is proclitic or enclitic. This can only be resolved by reference to morphological information. More generally, since our algorithm doesn't include direct knowledge of morphology (other than through knowledge of periphery), we will need to add this information if we are to match syllabification with morphology for words like "be+smirched", "be+stow", "bath+robes", and "birth+rates", which would be syllabified as */b ax s /-/ m er ch t /*, */b ax s / t ow /*, */ b ae th / r ow b z /* and */b er th /-/ r ey t s /*, respectively, by rule of sonority.

Sound	Sonority Index	Sound	Sonority Index
a	10	e,o	9
i,u	8	r	7
l	6	m,n	5
s	4	v,z,th(voiced)	3
f,th(voiceless	2	b,d,g	1
p,t,k	0.5		

Table 1: Sonority scale for phonemes.

Once the syllables are marked, we define the following three constituents as detailed in [2], where we distinguish between Enclitic and Proclitic in the Periphery.

Proclitic: Syllable component that only occurs morpheme initially. /s/ in (*still*) or /S/ in (*shrugged*).

Core: Syllable component common to all languages types. It contains the obligatory vowel.

Enclitic: Syllable component that only occurs morpheme finally.

These three parts, thus defined, capture a certain syllable structure, where P, $C1$, $C2$, and E (Figure 1) denote allowed sets of consonants, V denotes the set of vowels, and F denotes either a consonant or vowel, the latter being the second moraic element of a long vowel or diphthong. Given a word then, which is marked at the syllable level, it is possible to automatically find the three constituents. In a complex onset (consisting of more than one consonant), the first phoneme is marked as proclitic if it is */s/* or */S/.* In the Rhyme, consonants are marked as enclitic unless they are either an */s/*, an */l/* or an "assimilating nasal" occurring immediately after a short vowel. Assimilating nasals occur in words such as pump, rant, rank, combat, bandage, languid, ranch, hinge, mince, lens, triumph, etc. The "assimilating nasal" refers to a nasal consonant whose place of articulation (labial, laminal/apical-dentalveolar/ postalveolar, dorso-velar-lips, front-tongue, back-tongue), coincides with the place of articulation of the following consonant. Given these rules, we have therefore described the algorithm for marking core and periphery of syllables. The next step is then to syllabify a pronunciation dictionary so that core and periphery can be marked.

2.3. Evaluation

There is no validated reference syllabification by which to judge lexicon syllabification. So, in order to evaluate our algorithm, we want to syllabify a dictionary, which is already marked at the syllable level. The dictionary we are using for comparison has been developed at the Johns Hopkins summer school [5] and is a close variation of the high quality Pronlex lexicon, which has been automatically marked at the syllable level using Daniel Kahn's [4] Principle of English syllabification. Here, syllabification was controlled by three user-supplied lists: permitted syllable-initial consonant clusters (onsets), permitted syllable-final consonant clusters (codas), and prohibited onsets. This process is first run on native onsets and codas and then repeated for all words that failed syllabification by using corresponding lists of foreign onsets and codas while handchecking for satisfactory results. This syllabification algorithm used the generally accepted syllabification method that maximises onsets, assigning as many consonants as possible to syllable onsets while subject to the constraints of the list of permitted onsets. The dictionary contains around 71000 entries where we agreed on all but ca. 1300 syllabifications. In many cases, the phoneme */s/* was at the onset of a syllable in the dictionary while we assign */s/* to the coda (F or E) in certain compound words. Since conventional methods use beginnings of words as the way to model how syllables start, */thr/* in bathrobe, is allowed because it occurs in words such as 'throng'. English has the sequence */str/* at the beginning of words like "string", so that syllabification of "mistreat" for example is analysed as */mI/+/stri:t/.* Similarly, since English doesn't have short vowels at the end of words, in some models 'attitude' is analysed as */At/+/It/+/u:d/* rather than */A/+/tI/+/tu:d/* as in our algorithm. Such models often designate single consonants between vowels as "ambisyllabic"—ambiguous or belonging to both syllables).

Generally our syllable boundaries were correctly placed at the morphological boundaries more often than in the reference dictionary which can be explained with our indirect knowledge of morphology due to the knowledge of periphery. We take what happens at the beginnings and the ends of words to be exceptional, not the norm. We take syllable boundaries in the middle of words to be the way to model how syllables end and start generally. In addition, we differentiate between syllable transitions that occur where two morphemes meet and those that occur within a single morpheme. Though we can capture many morphologically correct syllables by this method, we need to extend our algorithm to include morphological knowledge in order to deal more effectively with prefixes and suffixes in the syllabification of words like "besmirch" */b ax s / m er ch/.*

3. FOREIGN ACCENT IDENTIFICATION

We expect speakers with greatest syllable structure differences between native and foreign language to have greatest difficulty with pronunciation in the foreign language. Similar to the example of the German accent, the behaviour of substitution of phonemes can be radically different for Core and Periphery of the syllable. We hypothesise a typology of syllable types based on Core vs. Periphery functions. At one end is English (or German) and at the other, tone languages like Vietnamese, Cantonese, Mandarin. Between these two extremes are languages without lexical tone with segmental configurations simpler than English. Syllable structures in tone languages tend to be comparatively simple in terms of phone segments, but are complicated by tones, each of which extends for the duration of a syllable or syllables expressing a grammatical unit, usually the word. The tone thus indicates the extent of the word. This difference in language typology has a strong effect on the ability to pronounce English in parts of the syllable that demarcate grammatical units. In order to study the structure of this type of foreign accent in English, we chose Vietnamese speech data. In contrast, Lebanese Arabic syllable structure has much more in common with English. We hypothesise that the pronunciation of English by Lebanese foreign speakers will be much closer to that of native speakers, and the variability less than that of a Vietnamese speaker.

3.1. DATA

The data used in this study come from the The Australian National Database of Spoken Language (ANDOSL [1]) [6]. The speech was recorded in an Anechoic chamber at the National Acoustics Laboratories of Sydney, Australia. We compare native Australian English to Vietnamese- and Lebanese-accented Australian English. The training set and test set for Australian English consist of one male speaker each. Each speaker read 200 phonetically rich and balanced sentences containing all types of phoneme combinations of Australian English pronunciation. Because the 200 sentences demanded a high degree of literacy from speakers for whom English was a non-native language, 50 sentences were chosen from the 200 and adjusted to have one member of every phoneme class in every permissible position. These were then read by the Vietnamese- and Lebanese-accented speakers. For Vietnamese, the training set and test set consist of six and three speakers respectively; the Lebanese training and test set consist of three speakers each. In order to analyze the accents, all speech was labelled by linguists with the closest Australian English phonemes achieved by the speakers. The second level of labeling consists of the transcribed words. Also available were a small dictionary covering all the words in the sentences that were uttered. This dictionary contained a single pronunciation model for each word representing the "ideal" speaker. Our syllabifier performs at 100% accuracy according to this dictionary which was syllabified by linguists.

Word	Syllable structure	actual pronunciation
1. The	D@(C)	/d/@:/
2. length	lE(C)NT(E)	/l/E/N/
3. of	O(C)v(E)	/O/b/
4. her	h@:(C)	/h/@:/
5. skirt	s(P)k@:(C)t(E)	/s/k/@:/s/
6. caused	ko:(C)zd(E)	/k/@u/s/
7. the	D@(C)	/d/@/
8. passers-by	pa:(C)s@(C)z(E)bai(C)	/p/a:/s/b/ai/
9. to	tu:(C)	/t/u:/
10. stare	s(P)te:(C)	/s/t/e:/

Table 2: Examples of English words as pronounced by a Vietnamese speaker. (E) denotes the Enclitic part, (C) the core part. Types of mistakes include: D→ d (1,7), deletion (2,8), Enclitic substitution (3,5), Enclitic devoicing (6), Enclitic simplification (6)

3.2. Aligning Utterances to Target Pronunciation

In order to study the accented speech as a function of syllable position, it is necessary to align the achieved phoneme sequence (handlabeled with English phonemes by linguists) with the target phoneme strings. An example sentence, in Table 2, "The length of her skirt caused the passers-by to stare" shows both target phonemes (in Australian English) and achieved phoneme string (as spoken by a sample Vietnamese speaker). The example shows how difficult it can be to align the two strings correctly in order to tag the syllable position of each of the actual pronunciations.

In the absence of a confusion matrix which could be obtained from training a phoneme recognizer, we use Dynamic Time Warping (DTW) in order to align the two strings with linguistic knowledge. The score to be maximized by matching achieved and target phoneme is calculated by summing up points as given in Table 3 over all shared categories over all possible phoneme pairs to be matched. Points listed in this table approximately reflect the degree of relatedness between two phonemes containing this feature. If we were to make a tree of all phoneme features, then the number reflects the depth of the tree at which is located a particular feature. For example, phonemes can be either vowels or consonants (1 point), vowels can be short or long (1.5 points), short vowels can be back or front (2 points). From this basic method, ambiguities are resolved with linguistic knowledge and points are altered by looking at the relative similarity of phonemes at different depths in the tree. So, for example, high short vowels and mid short vowels only receive 1 point, even at the same depth in the tree as back and front vowel. Matching /D/ (*loath*) to target /T/ (*bath*) results in a score: 1 (consonants) + 2 (fricatives) + 4 (laminodentals) + 1.5 (continuants) = 8.5. A perfect match to /T/ would have included 1.5 (voiceless). Matching /t/ to /T/, the score would result in 1 (consonants) + 2.5 (distal voiceless) + 1.5 (voiceless) = 5, which is smaller than 8.5; a less valuable match.

Category	Points	Category	Points
VOWELS	1	SHORT	1.5
LONG	1.5	BACK SHORT	2
CENTRAL SHORT	2	FRONT SHORT	2
BACKISH LONG	2	CENTRAL LONG	2
FRONT LONG	2	HIGH SHORT	1
LOW SHORT	1.5	MID SHORT	1
HIGH LONG	1	LOW LONG	1.5
MID LONG	1	DIPHTHONGS	1.5
RISING DIPH	3	FRONTING DIPH	0
CLOSING DIPH	3	CENTERING DIPH	2.5
INIT ROUNDING	1.5	FINAL ROUNDING	2
CONSONANTS	1	VOICELESS	1.5
VOICED	1.5	NASAL	4
LIQUID	4	APPROXIMANT	4
GLIDE	4	SONORANT	3
STOP	2.5	CONTINUANT	1.5
FRICATIVE	2	AFFRICATE	2.5
STOP FRIC	3	OBSTRUENT	1
LABIAL	2	LABIO DENTAL	4
LAMINO DENTAL	4	APICO ALVEOLAR	2
LAMINO POSTALVEOLAR	3	DORSO VELAR	4
DISTAL VOICELESS	2.5	DISTAL VOICED	2.5

Table 3: Linguistic Categories with corresponding points directly proportional to acoustic closeness (proportionate to number of common linguistic features).

The dynamic time warp returns two phoneme strings of the same length N, with each position, i, either marking a substitution, an insertion or a deletion. We thus have achieved an automatic method for marking the syllable

[1] More information on this database can be obtained at http://andosl.anu.edu.au:80/andosl/

position (Proclitic, Core, or Enclitic) within a pronunciation as inherited by the target dictionary pronunciation. While this method of alignment seems to work fine by inspection, it may be possible to improve the algorithm by acoustic analysis of closeness of phonemes within different categories.

3.3. Feature Analysis

Our goal is to look at the discrimination capability of features as a function of their position in the syllable. We want to see if position information improves the discrimination. Features used here correspond to occurrence frequencies of phoneme labels in the hand-labeled data for Vietnamese, Lebanese and Australian accented English. In order to identify discriminating features for any two classes of accented English speakers, it is essential to have a good estimate discrimination error due to a given feature. The estimate of the discriminability of two accents can be quantified for each feature based on a model of the feature distribution in the two accent classes introduced. We model each features by using a normal distribution, as shown in Figure 2, taking into account the mean occurrence frequency of a given feature, and the variation across speakers. Using this model, discriminating features can be extracted by estimating the Bayes' error due to two class-dependent distributions.

$$\text{Distance Measure} = \frac{1}{2}\exp -\frac{1}{4}\frac{(u_1[j]-u_2[j])^2}{s_1[j]^2+s_2[j]^2} \qquad (1)$$

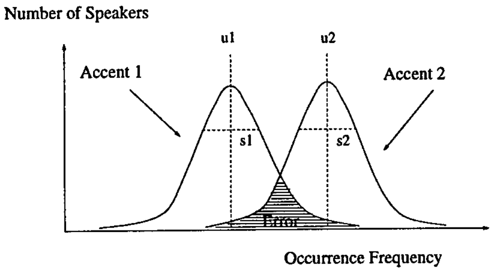


Figure 2: Normal Distribution.

For each of the features the corresponding discrimination error is estimated and thus we are able to look at the most important N features which will indicate the performance of accent discrimination based on this type of phoneme-based feature. Based on this model, we can now identify and sort the features by their classification error. Figure 3 depicts a graph of the top 40 features with respect to their corresponding estimated discrimination ability. From this graph, we can see that (1) Lebanese has less discriminating features which show less improvement when including position information. Vietnamese is a tone language and therefore, as expected, we see more improvement with this type of feature set.

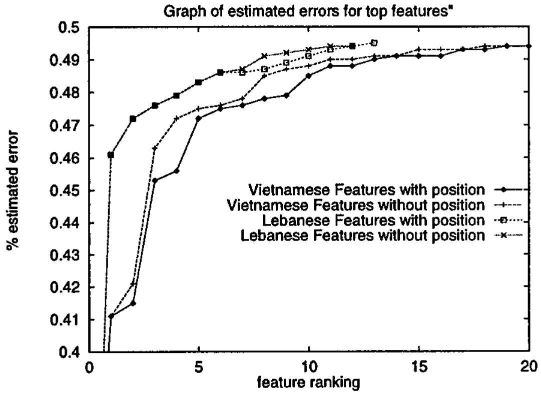


Figure 3: Top features or Lebanese vs. English and Vietnamese vs. English plotted as function of their estimated error and comparing position dependent features, with position independent Features. As expected, more improvement is seen in the Vietnamese list.

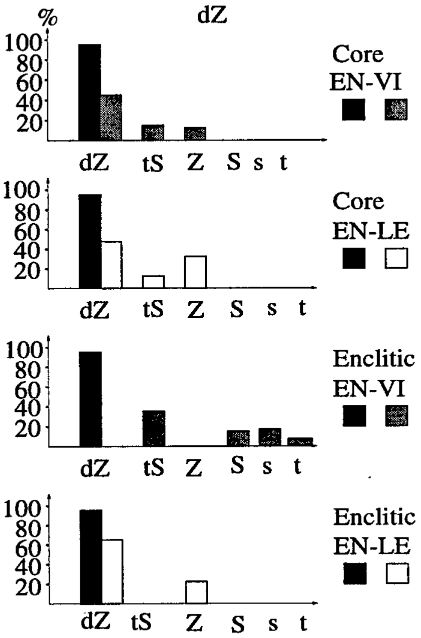


Figure 4: Comparison of language- and position-dependent substitutions for phonemes of /dZ/. Substitutions are different for Lebanese and Vietnamese and Core and Enclitic. Lebanese has less variability than Vietnamese.

3.4. Results

The total number of confusions is too large to describe here. In general, looking only at consonants, we can note the following trends:

- Confusions are different across accent groups.

- Confusions differ for *Enclitic* and *Core*.

- Lebanese speakers are more consistent in their substitutions than Vietnamese speakers. (See example for /dZ/ in Figure 4).

- Vietnamese accented speakers have a stronger accent than Lebanese accented speakers in terms of changes in voicing, manner, place and class. (See example for /dZ/ in Figure 4).

- The variability of the confusions is generally higher in the *Enclitic* than in the *Core* part of the syllable for both Vietnamese and Lebanese for /N/(*laughing*) and voiced fricatives.

- The variability of the confusions in the *Enclitic* is generally higher in Vietnamese than in Lebanese for stops, unvoiced fricatives, /T/, and /D/.

- phonemes /T/, /D/, /S/ and /z/(*zap*) are difficult for Vietnamese regardless of position.

- Voiced affricates are difficult for both accent groups.

- These trends are upheld across all speakers, however, the confusion probabilities vary.

One example, in particular, relates to the phoneme /d/ in Vietnamese. This phoneme is much more interesting for discriminability when treated as a function of position. In the Enclitic part its frequency is higher in English, but in the Core part its frequency is higher in Vietnamese. We now have the ability to study why this phenomenon takes place and why syllable position is so important. Table 4 lists some of the relevant confusions. We can see that /d/ is a substitute for /D/ (as 'th' in "the") for Vietnamese speakers—only in the Core part. In the Enclitic part of the syllable the pattern is quite different in that /D/ is simply devoiced. In addition, it can be seen that while /d/ is mostly pronounced correctly by Vietnamese speakers in the Core, /d/ is devoiced to /t/ in the Enclitic. All these effects combine to result in Vietnamese accent with a higher frequency of /d/ in the Core and a lower frequency of /d/ in the Enclitic when compared to native English.

Confusions including /d/

Position	Target	Achieved	English	Vietnamese
Core	D	D	0.99	0.33
	D	d	0.00	0.60
Enclitic	D	D	1.00	0.15
	D	T	0.00	0.27
	D	s	0.00	0.19
	D	t	0.00	0.27
Core	d	d	0.96	0.93
Enclitic	d	d	0.99	0.48
	d	s	0.00	0.12
	d	t	0.01	0.28

Table 4: Shows importance of location information of phoneme /d/ in Vietnamese accent.

3.5. Conclusions

No statistical analysis of these trends have been made due to the small amount of data used for analysis. However, having applied this information to a larger system, we have shown in [1] that accent identification can be improved by using syllable dependent information. In this paper we have shown that the position within the syllable is important because the pronunciation patterns of accented speakers vary as a function of the phoneme's position within the syllable and that the linguistic theory is reflected in real speech data and can be systematically captured. The linguistic understanding of this theory provides a means of predicting the discrimination potential for a given accent group when using this method. Having shown the connection between linguistics, theory and real data, we have gained the ability to reason about system performance at the linguistic level. This algorithm may also serve as a powerful tool for language teaching or alternatively for speaker identification/verification as certain habits of speakers might be captured much more effectively within the syllable constituents.

4. REFERENCES

[1] K. M. Berkling, Chris Cleirigh, Julie Vonwiller, and Marc Zissman. Improving accent identification through knowledge of english syllable structure. In *Proceedings International Conference on Spoken Language Processing*, Sydney, Australia, 1998.

[2] C. Cleirigh and J. Vonwiller. Accent identification with a view to assisting recognition. In *Proceedings International Conference on Spoken Language Processing*, volume 1, pages 375–379, Yokohama, Japan, apr 1994.

[3] J. A. Goldsmith. *Autosegmental and Metrical Phonology*. Basil Blackwell, 1 edition, 1990.

[4] Daniel Kahn. *Syllable-based generalizations in English phonology*. PhD thesis, Massachusetts Insitute of Technology, 1980.

[5] M. Ostendorf, B. Byrne, M. Bacchian, M. Finke, A. Gunwardana, K. Ross, S. Roweis, E. Shirberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfield. Modeling systematic variations in pronunciation via language-dependent hidden speaking mode. In *Proc. 1996 Summer Workshop on Speech Recognition*, 1996.

[6] J. Vonwiller, I. Rogers, Ch. Cleirigh, and W. Lewis. Speaker and material selection for the australian national database fo spoken languages. *Journal of Quantitative Linguistics*, 2(3):177–211, 1995.

Report of the plenary discussion on "Identification"

Chairperson: Louis Boves (KUN, the Netherlands)
Reporter: Sander van Wijngaarden (TNO-HFRI, the Netherlands)

Question by *Vloeberghs*: To what degree are speaker identification methods language dependent?

Van Leeuwen points out that the main problem in finding out the language-dependency of speaker identification methods appears to be the availability of sufficiently large multi-lingual corpora. Ideally, true multi-lingual speakers should be used to be able to carry out language-dependency tests for speaker recognisers. It is observed that such speakers are hard to find.

Boves suggests that another approach to answer the language-dependency question is to find theoretical considerations why some speaker identification methods should be less language-dependent than others. The method by Bimbot based on second order statistics, for instance, is found to be relatively language independent.

Schultz points out that with higher level methods (more linguistics involved instead of just speech acoustics), the more language dependency is to be expected.

Question by *van Wijngaarden*: How mature is language ID compared to other developments in the field of Speech technology?

Adda-Decker replies that Language ID is still very much in development. *Schultz* points out that, although for technologically less challenging applications (along the lines of MIVA) language ID may be at a workable level, more advanced applications will require considerable refinement of language identification methods.

Van Leeuwen asks if 'human benchmarks', the relative performance compared to humans, are available to use as a measure of maturity. Human benchmarks are available for most types of speech technology applications. *Berkling* replies that some human benchmarks are available for speaker and language ID. *Boves* remarks that it is difficult to obtain such benchmarks, and that human and machine performance are realised in different ways, and are hence difficult to compare. Furthermore, *Reynolds* remarks that human benchmarks are not very relevant in the case of identification; speaker and language ID systems are not used to replace humans; in fact, most applications will require super-human performance (for instance in terms of the number of languages/speakers that can be recognized simultaneously).

Question by *Boves*: How do we know if new information added to identification models is independent information, that will really add something to the model?

Berkling points out that if new information improves performance, one can generally conclude that it must have been useful and independent information. *Boves* remarks that the difficulty is predicting what information will be useful in advance, using a modeling approach rather than trial-and-error. For instance, one might wonder to which extent useful information for language identification is found in the consonants. *Marta* answers that empirical data suggests that, although consonants contribute largely to inter-speaker variability, there is little language information in consonants. *Van Leeuwen* remarks that this observation is somewhat contradictory to the fact that voice-onset times are known to discriminate strongly between some languages. *Boves* concludes that study of such phenomena may lead to a more model-driven approach to choosing which information to include in identification models.

Question by *Boves*: When does including new information that really does not add much start hurting performance? Will the saying 'garbage in, garbage out' apply?

Adda-Decker replies that the less 'blind' the method is, the better it will be. It is better to add, for instance, linguistic constraints than to rely on statistics. *Berkling* adds that linguistics constraints provide checks to verify if the system makes sense. Blind statistics will allow researchers to trust on methods that have little relation to linguistic reality. Then, 'garbage in, garbage out' will apply.

Final Review

1. Interoperability; what does it really mean?

One of the questions at the beginning of the workshop was about the precise meaning of the term 'interoperability.' Several aspects of the concept were distinguished in the discussions, *viz.*,

- Between systems
- Between people
- Between people and systems
- For different tasks

For all different meanings of the word the need for some kind of standardisation was expressed. Specifically, standardisation of the overall architecture of systems, or at the very least the definition of commonly agreed Application Programming Interfaces, was identified as an issue that urgently needs further attention. In very concrete terms the need for a standardised set of phone symbols used in various automatic speech systems was mentioned. In this context reference was made to the International Phonetic Association (IPA), which has spent a lot of effort in defining language independent phonetic units and their attendant symbols. Unfortunately, the work of the IPA is not widely known in the speech technology community. Reference was also made to SAMPA, a European attempt to define phone sets for all European languages, including a mapping to the IPA system and computer readable codes for all phones.

Another field where there is a need for agreement and eventually standardisation is related to measurement procedures. It was pointed out that measures which characterise individual modules of spoken language systems (*e.g.*, Word Error Rate for the acoustic decoder; perplexity for the language model) fail to predict important aspects of the performance of complete systems under many normal operational conditions. Moreover, as the performance of the modules increases, relatively crude measures like error rates will lose much of their diagnostic value. The participants of the workshop agreed that there is an urgent need for measures of the performance of 'systems in interaction.'

2. Open Systems: what can they do for us?

Closely connected to the issue of standards is the question whether the speech technology community would profit from open, public domain software and databases. On this issue the opinions remained divided. Some participants stressed the advantage of having common source code, *e.g.*, to allow everybody to use the 'same' basic ASR system. Many experiments, *e.g.*, student assignments in universities or pilot experiments in industry would be much easier and faster to perform if common—and therefore well-understood—software would e available. Other participants in the discussion expressed the feeling that few labs ever made changes to the HTK source code, when that was still freely available. These persons pointed out that the code of even a 'simple' ASR system is far too complex to allow for quick adaptations to the need of specific experiments. Moreover, mutual comparability is probably better supported by software systems that come in the form of executable code. Finally, the difficulties with maintaining commonality if all labs can change source code was mentioned. If the discussion had continued, agreement might have been reached on the statement that well documented public domain software will certainly facilitate experiments with the application of speech technology, but that this software does not need to be available in source code. Algorithmic research, on the other hand, will very likely require proprietary source code, that can only be understood, changed and maintained by a small number of specialist in a research lab.

Unlike the obvious lack of agreement about common software the opinion that the R&D community is crucially dependent on the availability of common databases for training and evaluation of modules and systems was unanimously supported. Yet, it was pointed out that there are many unresolved issues in this field. To begin with, there is a lack of multi-lingual databases; it is far from evident that the necessary multi-lingual databases can be designed and built with the limited resources that are presently available. In this context it was mentioned that there may be significant differences between SpeechDat style corpora (that have successfully be designed and collected for many languages over the last couple of years and which are now appearing in the catalogue of ELRA) and the corpora that are needed to develop multi-lingual dictation applications (which do not seem to be available through LDC or ELRA).

A short discussion addressed the question whether there is a need and a use for conversational databases in R&D in the field of Interoperability. Before this question can be decided more research is needed to better understand whether research on conversational databases does allow generalisations across the topic of the conversation and the tasks or assignments of the participants in the conversation.

3. Speech Science and Technology

In the field of multi-lingual operation and interoperability one cannot hope to obtain databases that are large enough to solve all modelling problems by straightforward statistics. As scarcity of data becomes more of an issue, the need for knowledge based approaches naturally increases. There was general agreement about the need for better integration of rule based and statistical approaches and for improved information exchange between technology and linguistics, but there were no ready made proposals for how to accomplish this. This stimulated quite some discussion about possible ways in which phonetic and linguistic knowledge can be brought to bear on the solution of speech technology problems. The discussion concentrated on questions about the relation between models and phonetic segments. Are these relations similar for HMMs and ANNs? No clear answers were given.

It was pointed out that conventional phonetic wisdom has its own inherent limitations. For instance, all phonetic and phonological theories seem to make a dichotomous distinction between consonants and vowels. In the speech signal such clear distinctions are seldom evident. Thus, there seems to be a need for phonetics and phonology to adapt their theories to better match the acoustic reality. In this context the work on trajectory modelling in technology and the corresponding research in non-linear phonology were mentioned as possibilities for bridging the gap. Most probably, phonetics and phonology can profit from the adoption of the procedures for analysing very large amounts of data that have been developed in speech technology. A clever combination of a data driven and a rule based approach should help to come to grips with the enormous range of speaker and language/dialect induced variation in the articulation and the acoustics of speech. In this context it is worth mentioning that the papers presented in the workshop showed contradictory results for the performance of data driven and rule based approaches to the clustering of models and states within models, necessitated by the lack of data. Further research is needed to learn whether the contradictions are in some sense fundamental, or whether they are rather artefacts caused by the specific tasks and databases used in the experiments.

Another part of the discussion addressed questions related to tone languages. The impression left by the discussion is that a lot of research remains to be done to develop proper ways of handling tone phenomena in ASR.

Another issue that was discussed, also without yielding a clear solution, was related to the importance of proper modelling of prosody in 'western' languages. It is clear that proper prosody is absolutely essential in text-to-speech synthesis. At the same time it may be an understatement to say that the potential contribution of prosodic parsing to speech recognition is much less obvious.

4. Open Issues

A number of questions were raised for which the answer is completely open. For one thing, the very basic question was raised how to define a language. When it comes to formal modelling the concept of a countable set of spoken languages may eventually prove to be wanting. For many foreigners as well as for most—if not all—present day ASR systems the difference between completely acceptable pronunciation by a person raised in Scotland and a person raised in East Anglia is probably as large as the difference between native and many types of non-native English. Several developers of ASR systems have made the experience that adding a few number of regional phones to the inventory of sub-word units improves recognition rate for speakers from that region very considerable (without compromising the performance for speakers from other regions).

The question of the 'best' units to model in phonetic theory, human perception as well as ASR was also discussed. Obviously, humans neglect an enormous amount of variation when they perceive and understand spoken language. Yet, not all of that variation is random, as testified by perception experiments with stylised versions of natural utterances. Maybe there is no single optimal set of units. It is very well possible that humans use units at the level of words, syllables, sounds and sub-phonemic phenomena in parallel, weighing those that seem to contribute most to the decoding problem more heavily than the others.

In multi-lingual speech technology research there is a clear need for detailed phonetic transcriptions of ever larger databases of accented and native speech. Quite naturally, this raised the question what level of accuracy can be obtained through auditory transcription, and what is the overall ratio between the effort (and therefore the money) invested in the transcription and the quality of the result. It was rightfully pointed out that the best measure that we presently have for transcription is actually based on agreement between multiple phoneticians who each transcribe the same passage. Yet, agreement cannot be simply equated to accuracy; it is quite possible that several phoneticians make the same 'mistake,' especially in circumstances where the verbal content of the message strongly suggests the presence/absence of a specific speech sound. Most participants who participated in the discussion expressed the opinion that even a high degree of agreement between experts may be difficult and expensive to obtain. This underlines the need for the development of powerful automatic tools to provide some form of phonetic transcription that is accurate enough for a range of application in multi-lingual R&D.

Some discussion time was also devoted to the issue of the relation between production and perception in cross-lingual problems. This is the well-known question whether learners of a second or foreign language are able to perceive phonemic and phonetic distinctions which they cannot produce, and—at least to some extent—vice versa. It is very difficult to design an experiment that provides an unequivocal answer to this question. Too much of the research reports in phonetics, language acquisition and second language learning is too strongly based on single subject studies, or on studies with small number of subjects with a common native language who take part in courses that teach the same new language. These experiments do not really allow the generalisation of the results to other subjects or other language pairs.

Limitations in the extent to which experimental results can be generalised to a wider range of contexts and situations were also discussed in connection to intelligibility tests that are completely based on the DRT or on other ways of using CVC utterances. There is a large body of experimental results suggesting that intelligibility in many languages, and especially in the languages of the Germanic and Slavonic families, is strongly related to consonant clusters. However, neither the DRT, nor any other test based on CVC stimuli, addresses consonant clusters. Given the potential ease of application of intelligibility tests with simple stimuli, it would be worthwhile to get a better understanding of the conditions under which 'real' intelligibility can be predicted from the results of DRT-like tests.

It was pointed out that the requirements we specify for multi-lingual ASR and Language ID systems are clearly super-human. Very few people are able to reliably identify more than some ten languages, and human performance in language ID drops considerably if the languages under investigation are not familiar. The same goes for speaker ID. Although humans may show superior performance in identifying persons from a relatively small group that they are highly familiar with, recognition performance suffers dramatically from memory limitations if humans are given the task to identify or verify speakers they are not familiar with. Last but not least, very few persons understand and speak, say, 15 languages, some of which may be totally unrelated to the family of the native language. Yet, we expect computers to perform these tasks at a high level of accuracy.

Another interesting discussion focused on the question about the relative importance of acoustic models, lexical representations and language models in ASR. It is very likely that part of the human advantage in speech recognition is due to the use of additional knowledge sources. Yet, it is difficult to imagine that the human advantage can be fully explained by pragmatic and/or linguistic intelligence. One example that suggest superior human performance in a task where additional knowledge sources are difficult to imagine is connected digit recognition (telephone numbers, credit card numbers, etc.). This kind of performance comparisons suggests that we still have a long way to go before machine performance is in the same league as human performance (as long as the tasks to be performed are within the range of what humans can do).

The last issue that deserves attention in this overview is related to so called Xenophones, *i.e.*, the phones that people produce if they pronounce foreign words in an utterance that for the rest is in their native language. The issue is of considerable importance in tasks that induce large proportions of foreign words in the input to a (human or automatic) system. Experiments with ASR systems indicate that recognition errors for foreign words are substantially more frequent than for phonetically similar native words. The problem is worst if the foreign words contain phonemes which do not occur in the native language, *i.e.*, xenophones. The realisation of xenophones shows a wide range of variation, certainly between speakers, but also within speakers. Productions can range from near-native (and therefore different from any model for native sounds), via some approximation of the non-native sound coloured by a neighbouring native sound, to substitution by a native sound. Modelling xenophones is extremely difficult and expensive, not in the last place because it requires detailed human phonetic transcriptions. Above it was already explained that such transcriptions are both expensive and error prone. The problem is only aggravated by the fact that, in read speech, it is difficult to predict what the impact of the spelling will be on the selection of the phones to be produced. One technique that is frequently used to collect xenophones is to ask subjects to read sentences containing names of foreigners. But many of these names can also be interpreted as native. If a subject fails to know the foreign person, she may well read the name as if it were native. Even if a reader recognises the fact that a name is foreign, she may not know the grapheme-to-phoneme correspondence in the target language; this will give rise to unpredictable selection of allophones or xenophones. For the time being no straightforward solution exists for the xenophone problem.

Lou Boves

Appendix – Author index

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	RTO-MP-28 AC/323(IST)TP/4	ISBN 92-837-1044-4	UNCLASSIFIED/ UNLIMITED

5. Originator
Research and Technology Organization
North Atlantic Treaty Organization
BP 25, 7 rue Ancelle, F-92201 Neuilly-sur-Seine Cedex, France

6. Title

Multi-Lingual Interoperability in Speech Technology

7. Presented at/sponsored by

the Information Systems Technology Panel (IST) Tutorial and Workshop held in Leusden, The Netherlands, 13-14 September 1999.

8. Author(s)/Editor(s)	9. Date
Multiple	August 2000

10. Author's/Editor's Address	11. Pages
Multiple	148

12. Distribution Statement There are no restrictions on the distribution of this document. Information about the availability of this and other RTO unclassified publications is given on the back cover.

13. Keywords/Descriptors

Speech recognition
Voice communication
Military operations
Speech
Human factors engineering
Linguistics
Intelligibility
Multilingualism
Interoperability

14. Abstract

Communications, command and control, intelligence, and training systems are more and more making use of speech technology components: i.e. speech coders, voice controlled C^2 systems, speaker and language recognition, and automated training suites.

Interoperability of these systems is not a simple standardisation problem as the speech of each individual user is an uncontrolled variable such as non-native speakers using, additional to their own language, an official NATO language. For international operations, this may cause a reduced performance or even cause malfunction of an action.

In order to address these topics a two-day workshop was organised focussed on the following subjects:

Non-native speech and regional accents
Cross language speech processing
Identification of language and speaker
Human Perception and Assessment.

This document presents the proceedings of the workshop and consists of twenty papers, four discussion reports and a final overview.

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25 • 7 RUE ANCELLE
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Télécopie 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DIFFUSION DES PUBLICATIONS RTO NON CLASSIFIEES

L'Organisation pour la recherche et la technologie de l'OTAN (RTO), détient un stock limité de certaines de ses publications récentes, ainsi que de celles de l'ancien AGARD (Groupe consultatif pour la recherche et les réalisations aérospatiales de l'OTAN). Celles-ci pourront éventuellement être obtenues sous forme de copie papier. Pour de plus amples renseignements concernant l'achat de ces ouvrages, adressez-vous par lettre ou par télécopie à l'adresse indiquée ci-dessus. Veuillez ne pas téléphoner.

Des exemplaires supplémentaires peuvent parfois être obtenus auprès des centres nationaux de distribution indiqués ci-dessous. Si vous souhaitez recevoir toutes les publications de la RTO, ou simplement celles qui concernent certains Panels, vous pouvez demander d'être inclus sur la liste d'envoi de l'un de ces centres.

Les publications de la RTO et de l'AGARD sont en vente auprès des agences de vente indiquées ci-dessous, sous forme de photocopie ou de microfiche. Certains originaux peuvent également être obtenus auprès de CASI.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Streitkräfteamt / Abteilung III
Fachinformationszentrum der Bundeswehr, (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

BELGIQUE
Coordinateur RTO - VSL/RTO
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
Directeur - Recherche et développement - Communications et gestion de l'information - DRDCGI 3
Ministère de la Défense nationale
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

ESPAGNE
INTA (RTO/AGARD Publications)
Carretera de Torrejón a Ajalvir, Pk.4
28850 Torrejón de Ardoz - Madrid

ETATS-UNIS
NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320

FRANCE
O.N.E.R.A. (ISP)
29, Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GRECE (Correspondant)
Hellenic Ministry of National Defence
Defence Industry Research & Technology General Directorate
Technological R&D Directorate
D.Soutsou 40, GR-11521, Athens

HONGRIE
Department for Scientific Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Centro documentazione tecnico-scientifica della Difesa
Via Marsala 104
00185 Roma

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

PAYS-BAS
NDRCC
DGM/DWOO
P.O. Box 20701
2500 ES Den Haag

POLOGNE
Chief of International Cooperation Division
Research & Development Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA - Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
VTÚL a PVO Praha /
Air Force Research Institute Prague
Národní informační středisko obranného výzkumu (NISČR)
Mladoboleslavská ul., 197 06 Praha 9

ROYAUME-UNI
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

TURQUIE
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar - Ankara

AGENCES DE VENTE

NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
Etats-Unis

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
Royaume-Uni

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Document Delivery
Montreal Road, Building M-55
Ottawa K1A 0S2, Canada

Les demandes de documents RTO ou AGARD doivent comporter la dénomination "RTO" ou "AGARD" selon le cas, suivie du numéro de série (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Des références bibliographiques complètes ainsi que des résumés des publications RTO et AGARD figurent dans les journaux suivants:

Scientific and Technical Aerospace Reports (STAR)
STAR peut être consulté en ligne au localisateur de ressources uniformes (URL) suivant:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR est édité par CASI dans le cadre du programme NASA d'information scientifique et technique (STI)
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
Etats-Unis

Government Reports Announcements & Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 2216
Etats-Unis
(accessible également en mode interactif dans la base de données bibliographiques en ligne du NTIS, et sur CD-ROM)

Imprimé par le Groupe Communication Canada Inc.
(membre de la Corporation St-Joseph)
45, boul. Sacré-Cœur, Hull (Québec), Canada K1A 0S7

NORTH ATLANTIC TREATY ORGANIZATION

RESEARCH AND TECHNOLOGY ORGANIZATION
BP 25 • 7 RUE ANCELLE
F-92201 NEUILLY-SUR-SEINE CEDEX • FRANCE
Telefax 0(1)55.61.22.99 • E-mail mailbox@rta.nato.int

DISTRIBUTION OF UNCLASSIFIED RTO PUBLICATIONS

NATO's Research and Technology Organization (RTO) holds limited quantities of some of its recent publications and those of the former AGARD (Advisory Group for Aerospace Research & Development of NATO), and these may be available for purchase in hard copy form. For more information, write or send a telefax to the address given above. **Please do not telephone.**

Further copies are sometimes available from the National Distribution Centres listed below. If you wish to receive all RTO publications, or just those relating to one or more specific RTO Panels, they may be willing to include you (or your organisation) in their distribution.

RTO and AGARD publications may be purchased from the Sales Agencies listed below, in photocopy or microfiche form. Original copies of some publications may be available from CASI.

NATIONAL DISTRIBUTION CENTRES

BELGIUM
Coordinateur RTO - VSL/RTO
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
Director Research & Development
Communications & Information
Management - DRDCIM 3
Dept of National Defence
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
VTÚL a PVO Praha /
Air Force Research Institute Prague
Národní informační středisko
obranného výzkumu (NISČR)
Mladoboleslavská ul., 197 06 Praha 9

DENMARK
Danish Defence Research
Establishment
Ryvangs Allé 1, P.O. Box 2715
DK-2100 Copenhagen Ø

FRANCE
O.N.E.R.A. (ISP)
29 Avenue de la Division Leclerc
BP 72, 92322 Châtillon Cedex

GERMANY
Streitkräfteamt / Abteilung III
Fachinformationszentrum der
Bundeswehr, (FIZBw)
Friedrich-Ebert-Allee 34
D-53113 Bonn

GREECE (Point of Contact)
Hellenic Ministry of National
Defence
Defence Industry Research &
Technology General Directorate
Technological R&D Directorate
D.Soutsou 40, GR-11521, Athens

HUNGARY
Department for Scientific
Analysis
Institute of Military Technology
Ministry of Defence
H-1525 Budapest P O Box 26

ICELAND
Director of Aviation
c/o Flugrad
Reykjavik

ITALY
Centro documentazione
tecnico-scientifica della Difesa
Via Marsala 104
00185 Roma

LUXEMBOURG
See Belgium

NETHERLANDS
NDRCC
DGM/DWOO
P.O. Box 20701
2500 ES Den Haag

NORWAY
Norwegian Defence Research
Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

POLAND
Chief of International Cooperation
Division
Research & Development
Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA - Centro de Documentação
Alfragide
P-2720 Amadora

SPAIN
INTA (RTO/AGARD Publications)
Carretera de Torrejón a Ajalvir, Pk.4
28850 Torrejón de Ardoz - Madrid

TURKEY
Millî Savunma Başkanliği (MSB)
ARGE Dairesi Başkanliği (MSB)
06650 Bakanliklar - Ankara

UNITED KINGDOM
Defence Research Information
Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

UNITED STATES
NASA Center for AeroSpace
Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320

SALES AGENCIES

NASA Center for AeroSpace Information (CASI)
Parkway Center
7121 Standard Drive
Hanover, MD 21076-1320
United States

The British Library Document Supply Centre
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom

Canada Institute for Scientific and Technical Information (CISTI)
National Research Council
Document Delivery
Montreal Road, Building M-55
Ottawa K1A 0S2, Canada

Requests for RTO or AGARD documents should include the word 'RTO' or 'AGARD', as appropriate, followed by the serial number (for example AGARD-AG-315). Collateral information such as title and publication date is desirable. Full bibliographical references and abstracts of RTO and AGARD publications are given in the following journals:

Scientific and Technical Aerospace Reports (STAR)
STAR is available on-line at the following uniform resource locator:
http://www.sti.nasa.gov/Pubs/star/Star.html
STAR is published by CASI for the NASA Scientific and Technical Information (STI) Program
STI Program Office, MS 157A
NASA Langley Research Center
Hampton, Virginia 23681-0001
United States

Government Reports Announcements & Index (GRA&I)
published by the National Technical Information Service
Springfield
Virginia 22161
United States
(also available online in the NTIS Bibliographic Database or on CD-ROM)

Printed by Canada Communication Group Inc.
(A St. Joseph Corporation Company)
45 Sacré-Cœur Blvd., Hull (Québec), Canada K1A 0S7

ISBN 92-837-1044-4