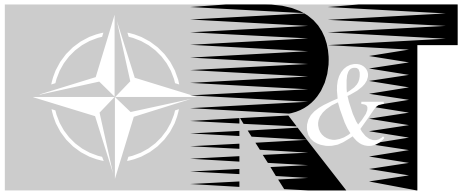
RTO-MP-066
AC/323(IST-025)TP/21

RTO-MP-066

NORTH ATLANTIC TREATY ORGANISATION



RESEARCH AND TECHNOLOGY ORGANISATION

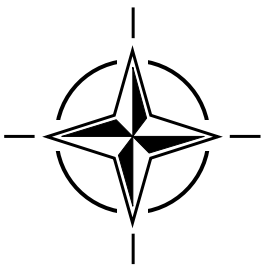
BP 25, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

RTO MEETING PROCEEDINGS 66

Multilingual Speech and Language Processing

(Le traitement multilingue de la parole et du langage)

Papers presented at the Information Systems Technology Panel (IST) Workshop held in Aalborg, Denmark, 8 September 2001.



Published April 2003

Distribution and Availability on Back Cover

This page has been deliberately left blank

Page intentionnellement blanche

RTO-MP-066
AC/323(IST-025)TP/21

NORTH ATLANTIC TREATY ORGANISATION

RESEARCH AND TECHNOLOGY ORGANISATION

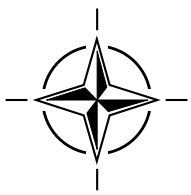
BP 25, F-92201 NEUILLY-SUR-SEINE CEDEX, FRANCE

RTO MEETING PROCEEDINGS 66

Multilingual Speech and Language Processing

(Le traitement multilingue de la parole et du langage)

Papers presented at the Information Systems Technology Panel (IST) Workshop held in Aalborg, Denmark, 8 September 2001.



The Research and Technology Organisation (RTO) of NATO

RTO is the single focus in NATO for Defence Research and Technology activities. Its mission is to conduct and promote co-operative 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 co-ordination 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 co-ordinates RTO's co-operation 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 co-operation.

The total spectrum of R&T activities is covered by the following 7 bodies:

- AVT Applied Vehicle Technology Panel
- HFM Human Factors and Medicine Panel
- IST Information Systems Technology Panel
- NMSG NATO Modelling and Simulation Group
- SAS Studies, Analysis and Simulation Panel
- SCI Systems Concepts and Integration Panel
- SET Sensors and Electronics Technology Panel

These bodies are made up of national representatives as well as generally recognised 'world class' scientists. They 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 co-operation 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.

Published April 2003

ISBN 92-837-1102-5

Multilingual Speech and Language Processing

(RTO MP-066 / IST-025)

Executive Summary

Multilingual speech and language technology is becoming recognized as an important issue for international organizations, both civilian and military. For instance, one might want to use a speech coder optimized for French in Germany or Turkey. A native speaker of Spanish might want to use a speech recognizer trained for American English. Additionally with the explosion of multilingual text material on the web, a British user might want to access Dutch documents using English search terms. For reasons such as these, a special task group of the NATO Research and Technology Organization (RTO) started a project on the development and assessment of multilingual speech and language applications.

To stimulate interaction between civil and military researchers and developers, the NATO Research Study Group on Speech and Language Technology (IST-011/RTG-001) organized a workshop in cooperation with the International Speech Communication Association (ISCA) in Aalborg, Denmark on 8 September 2001. Forty-eight researchers from 16 countries attended the workshop.

The workshop's focus was on the scientific and engineering aspects of speech and language processing for and despite multiple languages, dialects, non-native speech, and/or regional accents. A significant feature of the workshop was the announcement and discussion of a new database of native and non-native speech collected by the NATO Research Study Group. This database is called the NATO Native and Non-Native (N4) Speech Corpus, and it will be made available to researchers for further study.

The workshop opened with a keynote speech by Prof. Alex Waibel of Carnegie Mellon University, USA on multilinguality in speech and language systems. The keynote speech was followed by four technical sessions containing a total of 12 presentations and the workshop closed with a directed discussion session on various topics of interest to the workshop attendees. These proceedings contain the 12 technical papers presented at the workshop.

Le traitement multilingue de la parole et du langage

(RTO MP-066 / IST-025)

Synthèse

L'importance des technologies du traitement multilingue de la parole et du langage est de plus en plus reconnue par les organisations internationales civiles et militaires. Il se pourrait, par exemple, qu'un codeur vocal optimisé pour le français soit demandé en Allemagne ou en Turquie. De la même façon, un hispanophone pourrait avoir besoin d'un système de reconnaissance de la parole conçu pour de l'anglais américain. En outre, avec le foisonnement de textes multilingues affichés sur l'Internet, un utilisateur britannique peut souhaiter consulter des documents en néerlandais en se servant de termes de recherche en anglais. Pour de telles raisons, un groupe de travail de l'Organisation pour la recherche et la technologie de l'OTAN (RTO) a lancé un projet sur le développement et l'évaluation d'applications multilingues de traitement de la parole et du langage.

Dans le but de promouvoir des interactions entre personnels civils et militaires chargés de la recherche et du développement, le groupe d'étude OTAN pour la recherche IST-011/RTG-001 sur les technologies de la parole et du langage a organisé, en coopération avec l'Association internationale de la communication verbale (ISCA), un atelier, à Aalborg, au Danemark, le 8 septembre 2001. Quarante huit chercheurs de 16 pays différents y ont participé.

L'atelier a eu pour objectif d'examiner les aspects scientifiques et techniques du traitement de la parole et du langage du point de vue du multilinguisme, des dialectes, de l'expression non autochtone et/ou des accents régionaux. La présentation d'une nouvelle base de données de la parole autochtone et non autochtone constituée par le groupe d'étude OTAN pour la recherche et la discussion qui en a suivi a été l'un des moments forts de la manifestation. Cette base de données, appelée Corpus OTAN de la parole autochtone et non autochtone (N4), sera mise à la disposition de chercheurs aux fins d'études ultérieures.

L'atelier a débuté par un discours d'ouverture du Prof. Alex Waibel de l'université Carnegie Mellon (USA), sur le multilinguisme dans les systèmes de traitement de la parole et du langage. Ce discours a été suivi de 4 sessions techniques comprenant 12 présentations et l'atelier a été clôturé par une session de discussions dirigées sur différents sujets d'intérêt pour les participants. Le présent compte rendu contient les 12 communications techniques présentées lors de l'atelier.

Table of Contents

SESSION 4: LANGUAGE IDENTIFICATION AND MULTILINGUAL APPLICATIONS
Chairman: Marc ZISSMAN (US)

Foreword

The NATO Workshop on Multilingual Speech and Language Processing was held 8 September 2001 in Aalborg, Denmark at the Aalborg Kongres & Kultur Center, the site of EuroSpeech 2001. The workshop was sponsored by the NATO Research Study Group on Speech and Language Technology (IST-011/RTG-001) and supported by the International Speech Communication Association (ISCA). Although NATO sponsored the workshop, attendance was open to researchers from non-NATO countries. Of the 48 researchers from 16 countries attending the workshop, one quarter were affiliated with institutions from countries that are not members of NATO.

The workshop's focus was on the scientific and engineering aspects of speech and language processing for and despite multiple languages, dialects, non-native speech, and/or regional accents. A significant feature of the workshop was the announcement and discussion of a new database of native and non-native speech collected by the NATO Research Study Group. This database is called the NATO Native and Non-Native (N4) Speech Corpus, and it will be made available to researchers for further study.

The workshop opened with a keynote speech by Prof. Alex Waibel of Carnegie Mellon University, USA on multilinguality in speech and language systems. The keynote speech was followed by four technical sessions, and the workshop closed with a directed discussion session on various topics of interest to the workshop attendees.

The first session covered the topics of the N4 Speech Corpus and speaker recognition. In this session, Edouard Geoffrois (DGA/CTA/GIP, France) described the N4 Speech Corpus, and Marc Zissman (MIT Lincoln Laboratory, USA) presented results of speaker recognition experiments conducted on the N4 corpus. Joseph Campbell (MIT Lincoln Laboratory, USA) discussed the use of phonetic recognizers for multiple languages for speaker recognition.

The second session was on the topic of non-native speech. Sander van Wijngaarden (TNO Human Factors, The Netherlands) presented methods for the quantitative assessment of speech intelligibility in cross-language communication, and Carlos Teixeira (IST/INESC-ID, Portugal and SRI International, USA) discussed the evaluation of speakers' degree of nativeness using text-independent prosodic features. Laura Mayfield Tomokiyo (CMU, USA) discussed methods for adaptating speech recognition systems to handle non-native speech.

The third session covered speech recognition. Rathi Chengalvarayan (Lucent Technologies Inc., USA) presented an HMM-based speech recognizer for American, Australian, and British English. Andrej Žgank (University of Maribor, Slovenia) discussed crosslingual adaptation of multilingual triphone acoustic models, and Kåre Jensen (Nokia Mobile Phones, Denmark) discussed multilingual text-to-phoneme mapping for speaker independent name dialing in mobile terminals.

The fourth session was on language identification and multilingual applications. John Dines (Queensland University of Technology, Australia) discussed the fusion of output scores for language identification. Yasunari Obuchi (Hitachi, Ltd., Japan) discussed multilingual speech interpretation for cellular phones, and David Palmer (The MITRE Corporation, USA) discussed multilingual technology projects currently being undertaken in conjunction with the NATO BICES (Battlefield Information Collection and Exploitation) organization.

We would like to thank the NATO-RTO and ISCA for their support in the organization of the workshop, the speakers for their time, effort and expertise; the technical committee for their help in reviewing the proposals and their advice. Finely we would like to thank Prof. Børge Lindberg of the Center for PersonKommunikation at Aalborg University and Ms. Hanne Kristiansen of the Aalborg Tourist and Convention Bureau for the fine handling of the local arrangements.

Timothy R. Anderson and Raymond E. Slyh

Information Systems Technology Panel

CHAIRMAN

Dr Malcolm VANT
DRDC Ottawa
Defence R&D Canada-Ottawa
3701 Carling Avenue
Ottawa, Ontario K1A 0Z4
CANADA

VICE-CHAIRMAN

Dr René JACQUART
Directeur du DTIM
ONERA/CERT/DTIM
2, avenue Edouard Belin, BP 4025
31055 Toulouse, Cedex 4
FRANCE

IST-025/RWS-004 COMMITTEE MEMBERS

CHAIRMAN

Dr Timothy ANDERSON
Air Force Research Laboratory
AFRL/HECA
2255 H. Street
Wright Patterson AFB
OH, 45433-7022
UNITED STATES

MEMBERS:
Dr Stéphane PIGEON (Ecole Royale Militaire, BE)
Mr Carl SWAIL (Flight Research Lab/NRC, CA)
Dr Edouard GEOFFROIS (CTA/GIP, FR)
Prof. Jean-Paul HATON (Université Henri Poincaré, FR)
Dr Philippe VALERY (Sextant Avionique, IHM/ET, FR)
Ms Christine BRUCKNER (Bundessprachtenamt SM 1, GE)
Prof. Elmar NOETH (Universität Erlangen-Nurnberg, GE)
Prof. Carlos Jorge DA CONCEIÇÃO TEIXEIRA (INESC, PO)
Dr Herman J.M. STEENEKEN (TNO Human Factor Research Institute, NE)
Mr Ali ISKURT (TUBITAK-UEKAE, TU)
Mr Hasan PALAZ (TUBITAK-UEKAE, TU)
Mr J. Alan SOUTH (DERA Farnborough, UK)
Mr John J. GRIECO (AFRL/IFEC, US)
Prof. John HANSEN (University of Colorado at Boulder, US)
Dr Doug REYNOLDS (MIT Lincoln Laboratory, US)
Dr Clifford J. WEINSTEIN (MIT Lincoln Laboratory, US)

ORGANIZING COMMITTEE:
Timothy ANDERSON (Air Force Research Laboratory, US)
Raymond SLYH (Air Force Research Laboratory, US)

TECHNICAL COMMITTEE:
Herman J.M. STEENEKEN (TNO-Human Factors, NE)
Timothy ANDERSON (Air Force Research Laboratory, US)
Raymond SLYH (Air Force Research Laboratory, US)
Marc A. ZISSMAN (MIT Lincoln Laboratory, US)

PANEL EXECUTIVE

From Europe:

RTA-OTAN
Lt.Col. A. GOUAY, FAF
IST Executive
BP 25
F-92201 Neuilly-sur-Seine Cedex
FRANCE

From the USA or CANADA:

RTA-NATO
Attention: IST Executive
PSC 116
APO AE 09777

Telephone: +33 (1) 5561 2280 / 82 - Telefax: +33 (1) 5561 2298 / 99

Acknowledgements/Remerciements

The Panel wishes to express its thanks to the Danish RTB members for the invitation to hold this Workshop in Aalborg and for the facilities and personnel which made the Workshop possible.

Le Panel tient à remercier les membres du RTB du Danemark auprès de la RTA de leur invitation à tenir cette réunion à Aalborg, ainsi que pour les installations et le personnel mis à sa disposition.

This page has been deliberately left blank

Page intentionnellement blanche

The NATO Native and Non-Native (N4) Speech Corpus

Laurent Benarousse[1], *Edouard Geoffrois*[1], *John Grieco*[2], *Robert Series*[3], *Herman Steeneken*[4], *Hans Stumpf*[5], *Carl Swail*[6], *Dieter Thiel*[4]

[1]DGA/CTA/GIP, 16 bis avenue Prieur de la Côte d'Or, F-94114 Arcueil cedex, France
[2]AFRL/IFEC, 32 Brooks Rd., Rome NY 13441, USA
[3]20/20 Speech Ltd., MHSP Geraldine Rd., Malvern, Worcs. WR 14 3SZ, United Kingdom
[4]TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands
[5]Bundessprachenamt, Horbeller Strasse 52, 50354 Huerth, Germany
[6]Flight Research Laboratory, Building U-61, Montreal Rd., Ottawa Ontario, Canada
[7]ZU-StellenBwTAufkl, Kulmbacherst. 58-60, D-95032 Hof, Germany

Laurent.Benarousse@etca.fr, Edouard.Geoffrois@etca.fr, John.Grieco@rl.af.mil,
r.series@2020speech.com, steeneken@tm.tno.nl, hans.w.stumpf@t-online.de,
Carl.Swail@nrc.ca, Dieter.Thiel@bnhof.de

Abstract

The NATO Native and Non-Native (N4) corpus has been developed by the NATO research group on Speech and Language Technology, in order to provide a military-oriented database for multilingual and non-native speech processing studies.

Speech data has been recorded in the Naval transmission training centers of four countries (Germany, The Netherlands, UK and Canada). The material mainly consists in NATO English procedure between ships. In addition, the same speakers read a text ("The north wind and the sun") both in English and in their mother tongue.

The number of speakers per country ranges from 11 to 51, for a total of 115. The duration of speech ranges from 1.6h to 3.0h, for a total of around 9.5h.

This corpus can be used for various studies, including the influence of non-nativeness on speech, language and speaker recognition, and accent recognition.

1. Introduction

Speech technology is covering an increasing number of languages, and systems are becoming multilingual. They are also becoming more robust to speech variability such as speaking style and accents. However, for real applications, especially in a multilingual and multinational context, further robustness to regional and even non-native accents is necessary. Among the numerous corpora available for speech research, few have specifically addressed this issue. A workshop on multilingual interoperability of speech technology in 1999 [1] has shown that there is much interest in such issues, but much work remains to be done.

In this context, the NATO Speech and Language Technology group decided to create a corpus geared toward the study of non-native accents. In order to share a common, realistic and military-relevent task among the various member countries involved in the creation of the database, it was decided that the task would be naval communication, since it naturally includes much non-native speech, and because there are training facilities where data can be collected in several countries. The resulting corpus was called the NATO Native and Non-Native (N4) database, and is made available to the speech research community.

The corpus is described in detail in the following section. It can be used for various studies, and some preliminary experiments are described in section 3, before concluding.

2. The database

The database has been collected in four countries (Canada, Germany, The Netherlands and The United Kingdom) during naval communication training sessions. For each country, the main part of the recordings consists of NATO Naval procedure in English, where the typical sentence sounds like "This is alpha whiskey, roger. I make two seven zero six hostile, two seven zero six. Out." In addition, each speaker read a text ("The north wind and the sun") both in English and in his mother tongue (when different).

The audio material was recorded on DAT and downsampled to 16kHz-16bit, and all the audio files have been manually transcribed and annotated with speakers indentities using the tool Transcriber [2]. Navy procedure recordings and text reading have been stored in different files, and the first digit in the file name indicates the type of speech.

The duration of signal per country ranges from 3.2h to 6.3h for a total of around 20h, and the duration of speech ranges from 1.6h to 3.0h, for a total of around

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

9.5h[1]. Among speech segments, the durations of Navy procedures recordings range from 1.3h to 2.3h for a total of 7.5h, durations of text reading in mother tongue range from 1.5min to 22.9min, for a total of around 1h and in non-native English from 16.5min to 22.5min, for a total of around 1h. Table 1 summarizes these durations.

Table 1: Durations of the database in hours.

	CA	GE	NL	UK	All
Signal	**5.30**	**3.20**	**5.00**	**6.30**	**19.80**
→ Silence	3.00	0.56	2.00	4.70	10.26
→ Speech	2.30	2.64	3.00	1.60	9.54
Speech	**2.30**	**2.64**	**3.00**	**1.60**	**9.54**
→ Navy proc.	2.00	1.90	2.30	1.30	7.50
→ Read text	0.30	0.74	0.70	0.30	2.04
Read text	**0.30**	**0.74**	**0.70**	**0.30**	**2.04**
→ Non Native	0.27	0.37	0.32	0.00	0.96
→ Native	0.03	0.37	0.38	0.30	1.08

For each speaker much information is stored in the documentation of the database. Gender, age, weight, length, possible speaking or hearing disorders, education level, living area, accent, second language, smoking or non-smoking, where and when English was learnt (for non-native speakers).

Table 2: Speakers information.

	CA	GE	NL	UK	All
# Speakers	22	51	31	11	**115**
# Women	5	0	9	5	**19**
Age	22-35	17-23	17-61	19-62	**17-62**
Age mean	28.3	20.1	21	27.5	**22.6**

The speakers accents vary a lot among each country, and the number of speakers goes from 11 to 51, for a total of 115. The average age is 22.6 and the women represents 18% of the population. Table 2 summarizes the informations concerning the speakers.

3. Preliminary experiments

In this section, experiments that have been done on the N4 database are presented. Speaker recognition results are available in an other article [3], as LID studies are detailled below.

3.1. Speaker recognition

Three speaker recognition systems developed at TNO Human Factors, the US Air Force Research Laboratory, Information Directorate and MIT Lincoln Laboratory have been tested on segments from the Dutch part of the N4 database. During these preliminary experiments,

[1]We counted as speech any transcribed signal segment.

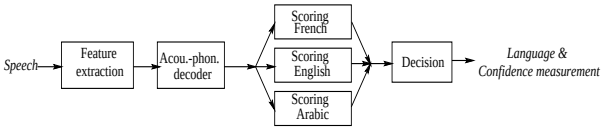


Figure 1: General LID system architecture using an acoustic-phonotactic approach

and among others studies, the authors investigated the impact of cross-language training and testing.

3.2. Language identification

3.2.1. Baseline system

A LID system that has been developed by CTA using LIMSI software [4] was tested on speech from the database. Only the native part (reading of "The north wind and the sun") of the recordings was used.

The baseline system uses a conventional acoustic-phonotactic approach for three languages, associated with a rejection strategy. One acoustic-phonetic decoder is used to perform a phonetic transcription of the speech, and each of phonotactic models returns a score for this transcription. The best score identifies the language. A rejection strategy then allows to give a confidence measurement with this decision (Figure 1).

The decoder uses a 3-state HMM with 32 gaussians per state with a set of 91 language-independent and context-independent phonetic units. This HMM has been trained on a large amount of automatic labeled Broadcast News recordings in three languages (17h per language). Phonotactic modules are trigrams models that have been trained of 17 hours of transcription from monolingual speech.

Because the original LID system has been developed for French, English and Arabic, only French and English phonotactic models were kept for the experiments. The 2-languages LID system was then tested on the French and English segments of the database.

3.2.2. Experiments

For LID systems segment duration is a key-parameter. Figure 2 provides results comparing performances of the system as a function of duration. Because of the number of segments is less important for French than for English (from 8.7% to 4.5% of the total number of segments are in French, depending of the duration), the global error rate follows more or less the English's one. French and English share the same kind of evolution of LID error rate as a function of duration (more than 88% when going from 2s segments to 20s segments). For a duration of 20 seconds, the system identifies the spoken language almost with no mistake (2.3% of error rate).

Since French from the database has a completely dif-

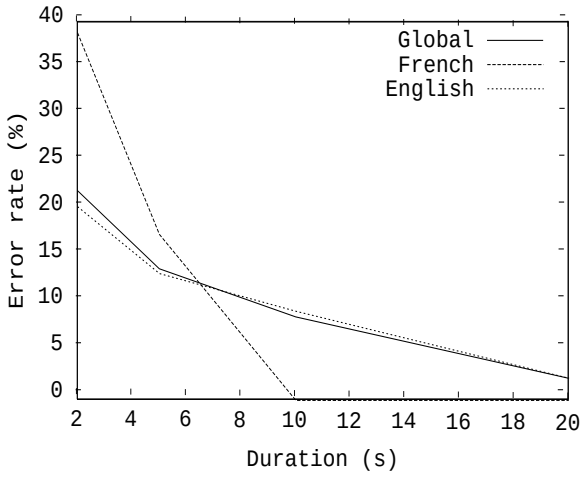


Figure 2: LID error rates as a function of test segments duration. Results are given for each language and for the global system.

ferent accent[2] than the French used for training the system, it seems that the LID system is rather robust toward accent, at least for French.

4. Conclusion

In this paper, we have presented the NATO Native and Non-Native corpus (N4). Because the data have been collected in four different countries during realistic communication training, in English and in mother tongue (for non-English speakers), the database can be used for various experiments. Speaker recognition and language identification, that have been presented in this article, are some examples of what can be done with this corpus.

5. References

[1] "Multi-lingual Interoperability in Speech Technology", Leusden, The Netherlands, September 1999.

[2] C. Barras, E. Geoffrois, Z. Wu, M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production", *Speech Communication*, Vol. 33, No 1-2, January 2000.

[3] M.A. Zissman et. al., "Preliminary speaker recognition experiments on the NATO N4 corpus", *Proceedings of Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, September 2001.

[4] L. Benarousse, E. Geoffrois, "Preliminary experiments on language identification using broadcast news recordings", *Eurospeech'01*, Aalborg, Denmark, September 2001.

[2]The speakers are from Quebec, where the French accent is very different than the one from France.

This page has been deliberately left blank

Page intentionnellement blanche

PRELIMINARY SPEAKER RECOGNITION EXPERIMENTS ON THE NATO N4 CORPUS

M.A. Zissman[1], R.A. van Buuren[2], J.J. Grieco[3], D.A. Reynolds[1], H.J.M. Steeneken[2], M.C. Huggins[4]

[1]MIT Lincoln Laboratory, Lexington, MA USA
[2]TNO Human Factors, Soesterberg, The Netherlands
[3]Air Force Research Laboratory, Rome, NY USA
[4]ACS Defense, Inc., Rome, NY USA

maz@ll.mit.edu, vanBuuren@tm.tno.nl, John.Grieco@rl.af.mil,
dreynolds@ll.mit.edu, steeneken@tm.tno.nl, hugginsm@rl.af.mil

ABSTRACT

The NATO N4 corpus contains speech collected at naval training schools within several NATO countries. The speech utterances comprising the corpus are short, tactical transmissions typical of NATO naval communications. In this paper, we report the results of some preliminary speaker recognition experiments on the N4 corpus. We compare the performance of three speaker recognition systems developed at TNO Human Factors, the US Air Force Research Laboratory, Information Directorate and MIT Lincoln Laboratory on the segment of N4 data collected in the Netherlands. Performance is reported as a function of both training and test data duration. We also investigate the impact of cross-language training and testing.

1. INTRODUCTION

From 1999-2001, the NATO Research Study Group on Speech and Language Technology (IST-011/TG-001) coordinated the collection of the NATO Native and Non-Native (N4) corpus [1]. This corpus contains speech collected at naval training schools within several NATO countries. The speech utterances comprising the corpus are primarily short, tactical transmissions spoken in English and typical of NATO naval communications.

The N4 corpus was intended to support research and development of automatic speech processing systems for tactical military applications characterized by limited vocabulary, short utterance length and non-native speakers. It is hoped that the corpus will be of use to developers of the following types of speech processing systems:

- Speech recognition – full transcription,
- Speech recognition – call-sign identification,
- Language understanding, and
- Speaker recognition.

Because many of the N4 speakers are non-native speakers of English, the corpus should be especially useful for both developing speech processing systems capable of processing non-native speech effectively and for evaluating the quantitative performance of such systems.

This paper reports on some preliminary speaker recognition experiments performed on the N4 corpus. The primary purpose of these experiments was to compare performance of a variety of existing speaker recognition systems on short utterances composed of tactical military speech. We also wished to show the impact of non-native, cross-language training/testing on speaker recognition performance. Finally, we hoped that by running and reporting the results of some initial speaker recognition experiments, we might encourage others to use the corpus for their own research purposes.

The rest of this paper is organized as follows. Section 2 describes in detail the parts of the N4 corpus that were used for our preliminary speaker recognition experiments. Section 3 describes the three speaker recognition systems that were run. Section 4 outlines the experiments and reports the results. In Section 5, we discuss the results and suggest possible future directions.

2. N4 DATA SELECTED FOR EXPERIMENTATION

Our preliminary experiments have used only a fraction of the available N4 data. Although data from four countries, Canada, Germany, the Netherlands and the UK are available in the corpus, time constraints forced us to limit our experiments to the Netherlands (NL) data. We chose to process the NL data because several of us (Van Buuren

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

and Steeneken) had been involved in its collection; hence, we were more familiar with this segment of the N4 corpus than with those segments collected in other countries. Additionally, though the German segment of N4 has more speakers, we thought the NL segment had an adequate number of speakers for these initial experiments.

The following sections describe the characteristics of the NL data including a description of the speakers, the types of data that were recorded, and how the data were used for these preliminary experiments.

2.1 The NL Speakers

The NL data were collected from 30 speakers over three days at the Operational School, Royal Dutch Navy at Den Helder, the Netherlands. The 21 males and 9 females were all between 17 and 25 years old at the time the data were collected. These speakers were all enlisted naval personnel receiving training in naval communications. All of the speakers were native speakers of Dutch. Of the 30 speakers, 19 were also fluent in English, and some were also fluent in German and Spanish.

2.2 The NL Tactical Speech

The first type of speech collected was tactical military communications. Although the speech was collected from personnel in the Royal Dutch Navy, all of the tactical speech was spoken exclusively in English. Each speaker wore a noise canceling microphone headset typical of the type used on Dutch ships. The microphone on the headset was connected to a digital audio tape recorder sampling at 48 kHz with 16-bit resolution. The speech data were later digitally low-pass filtered and down-sampled to 16 kHz for subsequent processing.

The speech collected from each speaker was recorded in a single session lasting between one and two hours. A total of three sessions were recorded, with 11 speakers recorded in the first session, 16 in the second session and 7 in the third session. Each speaker was assigned a ship name, call sign and network name. Speakers were given a list of actions to perform that required communicating from one "ship" to another. Within a session, speakers took turns performing these actions. Though speakers composed their messages spontaneously from the instructions (i.e. the speakers did not merely read from a script), the content of each message (i.e. choice of words, word order) was constrained by the specific action to be achieved and by NATO communication standards. Some example transcriptions from these transmissions are:

```
papa alfa zulu juliett this is papa alfa
sierra zulu reporting into net over
```

```
papa alfa sierra zulu this is papa alfa
mike alfa I authenticate quebec over
```

Once the exercises were complete, human labelers reviewed the speech collected and produced time-aligned speaker and orthographic transcriptions. These transcriptions were used to divide the session-long recordings into single-speaker speech fragments that have an average length of 5.9 seconds. The number of fragments produced per speaker ranged from 7 to 45 depending on the frequency with which that speaker spoke during the exercise.

For the purposes of these preliminary experiments, each tactical speech fragment for each speaker was assigned to either the training set or the test set. The baseline tactical speech *training set* consisted of the first five and last five fragments collected from each speaker. The tactical speech *test set* consisted of the remaining speech fragments. For four of the speakers, ten or fewer fragments were available; therefore, data for these four speakers were totally discarded leaving a total of 26 speakers for our experiments. There were a total of 260 fragments in the training set (26 speakers, 10 fragments per speaker) and 383 fragments in the test set (26 speakers, 15 fragments per speaker on average, with a minimum of 3 fragments and a maximum of 35 fragments).

Because each speaker participated in only a single recording session, the direct applicability of our experiments to real military applications is limited. The fact that the speaker's physical condition and many other recording characteristics were invariant during the single recording session results in an acoustic similarity between a speaker's training and testing data that is much closer in the NL data than we would expect in many (but not all) military applications.

2.3 The NL Read Speech

In addition to the tactical speech described above, the NL segment of the N4 corpus also contains some read speech for each of the 30 speakers. The same recording equipment and configuration were used for collecting the read speech as were used for collecting the tactical speech. Each speaker was asked to read a short parable in both English and Dutch. An example sentence from the English version of the parable is given below:

```
The North Wind and the Sun were disputing
which was the stronger when a traveler came
along wrapped in a warm cloak.
```

The average duration of the read English parable was 36 seconds. The average duration of the read Dutch parable was 34 seconds.

For the purposes of these preliminary experiments, the read speech was used as an alternative set of training data allowing us to compare cross-style training and testing (i.e. train on read speech, test on tactical speech) and cross-language training and testing (i.e. train on English read speech and test on English tactical speech; train on Dutch read speech and test on English tactical speech).

3. THE SPEAKER RECOGNITION SYSTEMS

Three speaker recognition systems were evaluated. Each of these systems operates in two phases. During the *training phase*, training speech from each of the speakers is used to create speaker models. The three systems use different forms of speaker models and use different algorithms to produce these models. During the *recognition phase*, test utterances are compared to the speaker models. This comparison process is different for each of the three systems, but it generally results in the production of one score per model, with higher (or lower) scores indicating a closer match. What follows below is a very brief description of each of these system. References to detailed descriptions of each of the systems are also provided.

3.1 The TNO System

The TNO automatic speaker recognition system is based on an algorithm by Bimbot *et al*. [2]. The algorithm takes as its inputs the covariance matrices of frame-based features (e.g., third-octave spectra), computed over an entire utterance: these covariance matrices can be regarded as templates. Output of the algorithm is a distance measure that is computed between templates in the training set and the template of the unknown (test) speaker. The training template for which the shortest distance is found corresponds to the most likely speaker of the test template.

The TNO system is used primarily for recognition of short and mostly low-quality speech (noise, low bandwidth). Computational complexity is relatively low, so it can easily be implemented on a standard PC and used in the field. It has been evaluated in the past on the TIMIT 420-speaker training set, on which error rates in a closed-set recognition experiment were well below 5% (which is in good agreement with the results by Bimbot). Using the system on a private radiotelephone database with eight speakers and fragment lengths primarily between 2 and 5 seconds, the error rate is about 15%. Furthermore, the system has proven to be largely language-independent.

In its present form, the algorithm is used in closed-set scenarios only. For the current experiments, we used two parameter sets (10^{th} order LPC coefficients and third-octave spectra) computed from the full 8 kHz bandwidth of the input speech. We computed a weighted summed distance from the individual distance measures for each of the parameter sets. To improve reliability, an N nearest-neighbor selection can be applied when evaluating the distance measures. In the current experiments, best results were obtained using N=1 (i.e. only the shortest distance is used for the selection of the most-likely speaker).

3.2 The AFRL System

The AFRL speaker identification system run on the NL data is based on the multi-feature classifier fusion technique described by Ricart *et. al.* [3]. During training three features were extracted from the speech and input to the Linde-Buzo-Gray vector quantization (VQ) training algorithm (also known as the Generalized Lloyd's algorithm) [4]. These three features were the cepstrum, delta-cepstrum and liftered cepstrum, all derived from linear prediction coding coefficients. The features were used to train an independent model for each speaker, with three separate VQ codebooks for each speaker. This approach is slightly different from the more conventional feature concatenation method in which only one codebook per speaker is generated. During testing the classification result is based on the L2 distance metric. The total distance score for each trained speaker's model is calculated from a sum of distances to unknown feature vectors. These codebook distances are normalized by the number of test vectors. The distances resulting from each feature classifier are then fused through an adjudication process. This process is a linear fusion technique in which the classification distances for each feature are normalized with respect to the minimum distance for that feature. The speaker with the lowest linearly fused sum is declared the winner in the closed-set case.

For the open-set experiments each speaker was trained as described above, but cohorts were used for scoring normalization. The entire training dataset was used as a pool of cohorts.

Most communication systems have 4kHz of bandwidth, so we ran some experiments that had the effect of artificially band limiting the data to 4kHz. Use of the AFRL Usable Data Bandwidth (UDB) algorithm [5] revealed the average usable speech bandwidth on the training set was 100 Hz to 7.7 kHz. This rather wide bandwidth was not unexpected because the data set was collected with high quality microphones and recording equipment. Therefore, we also ran experiments that used features capable of modeling the full bandwidth. In all these experiments, channel normalization was disabled.

3.3 The MITLL System

The MIT Lincoln Laboratory speaker recognition system uses the Gaussian Mixture Model (GMM) and Universal Background Model (UBM) approach developed by Reynolds [6]. Front-end processing converts sampled speech waveforms into mel-weighted cepstra and delta-cepstra at a 100 Hz frame-rate. In the original version of this system, the training speech from each speaker is used to produce a single GMM for that speaker using the Estimate-Maximize algorithm. The order of the GMM varies depending on the amount of training speech available. In the UBM version of this system, a single universal GMM is trained from a large number of non-target speakers. Models for individual speakers are created by adapting the UBM. During recognition, the likelihood of the test speech is computed for each of the GMMs produced during training. For closed-set recognition, the speaker corresponding to the most likely GMM is hypothesized as the speaker of the test utterance. For open-set speaker detection, likelihood ratios are computed on a per-test-message basis. These ratios can then be sorted to produce a list of test messages ordered by the likelihood that they were spoken by a specific speaker.

We used a 512-mixture GMM UBM system for all experiments on the NL data. The speaker training data from the baseline experiment was used to train the UBM. To allow the use of the same system for both closed-set identification and open-set detection, we trained 26 different UBMs, where each UBM excluded the training speech from one of the 26 speakers. For a test message from speaker 1, for example, the UBM that excluded speaker 1's training data was used. This was required so as not to violate the open-set assumption. In all NL data experiments, features were extracted from the entire 8 kHz bandwidth of the input speech, and channel normalization was disabled.

The MITLL system has been evaluated on unconstrained telephone speech in annual NIST evaluations in which it has exhibited top performance on a variety of detection tasks (e.g. single-speaker detection, two-speaker detection) [7]. Closed-set recognition experiments on 630 speakers of the TIMIT corpus have yielded error rates of less than 1% [8]. The test tokens in the TIMIT experiments were single sentences having durations of approximately 3 seconds.

4. EXPERIMENTS AND RESULTS

Our preliminary experiments were designed to measure the performance of the various speaker recognition systems as a function of amount of training data (i.e. number of training fragments per speaker), type of training data (tactical vs. read) and language of training data (English vs. Dutch). The sub-sections below describe the baseline experiment and several of the contrast experiments.

4.1 The Baseline Experiment

In the baseline experiment, each speaker model was trained on ten fragments of tactical training speech, i.e. the baseline experiment used all the speech in the training data set. The test set consisted of the 383 tactical test fragments. Baseline experiment closed-set results are shown in Table 1 below.

Duration	Number of tokens	TNO	AFRL 4kHz	AFRL 8kHz	MITLL
All	383	17%	3%	1%	1%
1-3 sec	132	28%	8%	3%	2%
3-6 sec	132	13%	2%	0%	0%
6-27 sec	119	8%	0%	0%	0%

Table 1. Results of the baseline experiment. Recognition results shown in terms of closed-set error rate.

The results shown in columns 3, 4, 5 and 6 of Table 1 indicate the error rate of each of the three systems on the 26-alternative, forced-choice speaker recognition experiment. The two AFRL results show performance of the narrowband and wideband results, respectively. The first row of results indicates performance on all 383 test fragments. The subsequent rows indicate performance on subsets of the test fragments having durations of 1-3 seconds, 3-6 seconds and 6-27 seconds. For the results on all 383 test fragments (the first row of results), the difference in performance between the TNO and other systems is statistically significant at the 95% confidence level. The differences in performance among the two AFRL and MITLL systems are not statistically significant at the 95% level. We also observe a general decrease in error rate as test-utterance duration increases.

4.2 Reducing the Amount of Training Data

A set of experiments was run to measure the impact of reducing the amount of training data available to produce the speaker models. Three contrasts were run:

- Train on the first three and last three training fragments only.
- Train on the first three training fragments only.
- Train on the single longest training fragment of the ten available in each speaker's training set.

For each of these contrasts, the test set was identical to that used in the baseline experiment. Table 2 shows the results of this set of experiments.

Training fragments Per spkr	Avg duration of training speech/spkr	TNO	AFRL 4kHz	AFRL 8kHz	MITLL
10 (baseline)	61.0 sec	17%	3%	1%	1%
6	36.3 sec	26%	10%	4%	2%
3	18.4 sec	44%	26%	15%	5%
1	13.7 sec	34%	36%	22%	9%

Table 2. Results of the reduced training experiments. Recognition results shown in terms of closed-set error rate.

We observe that, in general, error rate increases as the amount of training speech decreases.

4.3 Training on Read Speech

The final set of closed-set experiments examined the impact of training on read speech (either English or Dutch) while testing on the tactical test speech (English only). Table 3 shows the results of these experiments.

Training Material	Avg duration of train speech/spkr	TNO	AFRL 4kHz	AFRL 8kHz	MITLL
English Tactical	36.3 sec	26%	10%	4%	2%
English Read	36.0 sec	33%	21%	8%	8%
Dutch Read	34.2 sec	33%	18%	12%	9%

Table 3. Results of the training on read speech. Recognition results shown in terms of closed-set error rate.

We observe that performance of the three systems is somewhat degraded when training on read speech vs. tactical speech. The impact of the training/testing language mismatch is system dependent.

4.4 Detection Results

The results reported above show the ability of the various speaker recognition systems to perform closed-set identification. These results are relevant to situations where the set of possible speakers is known in advance and when training speech is available for all of these speakers. However, in applications such as speaker authentication, the set of speakers that could be encountered during recognition is unconstrained. In such cases, it can be more meaningful to report results of detection experiments. The general approach is to begin by designating a specific speaker as the target speaker. Test fragments are then sorted according to their score against the target speaker model. Generally, this score should not be a raw likelihood but should be a likelihood ratio or a posterior probability, i.e. it should be normalized against scores obtained for the same test fragment against other speaker models. Those test fragments with high likelihood ratios are assumed to be more likely to have been spoken by the target speaker than those with lower likelihood ratios. For any score threshold, we can compute the probability of false alarm, i.e. the number of test fragments above threshold that were not spoken by the target speaker, and the probability of miss, i.e. the number of test fragments below threshold that were spoken by the target speaker. We can repeat this single-speaker detection experiment for all speakers, and we can average the results.

The exact details of the normalization process are system dependent, but to preserve the open-set nature of the experiment, we impose one constraint:

> If speaker *i* is the target speaker, then when producing a score for test fragments spoken by speaker *j*, $j \neq i$, systems may not use the score from speaker *j*'s model for normalization purposes.

This constraint ensures a true open-set test because we are not exploiting any speaker *j* training data as we score speaker *j* background test fragments.

Both the AFRL and the MITLL system can produce normalized, sortable scores. Table 4 shows the equal-error rate (the point at which the probability of false alarm and probability of miss are equal) for those two systems for a few conditions.

Duration	AFRL 4kHz	AFRL 8kHz	MITLL
Baseline	5%	3%	1%
6 fragment training	8%	4%	2%
English Read Speech	10%	6%	7%

Table 4. Single-speaker detection results. Results are shown in terms of equal-error rate.

5. DISCUSSION

This paper has reported on our experience running speaker recognition experiments on the N4 corpus. Our major conclusions are:

- Given sufficient training data, high-performance speaker recognition can be obtained on the short tactical utterances from the NL data set.
- There were some statistically significant differences in performance among the speaker recognition algorithms that were evaluated.
- System performance is largely determined by the complexity of the model (e.g. number of parameters) employed, with simpler systems (e.g. TNO) having somewhat higher error rate than more complex systems (e.g. AFRL and MITLL).
- Reducing the duration of training data and testing data generally increased the speaker recognition error rate.
- Cross-style and cross-language training/testing had a system-dependent impact on error rate.

Paths for further speaker recognition experimentation on the N4 data set could include:

- Expansion of the experiments to include data from the three other countries represented in the N4 data set. We believe that all such experiments should be country-specific, however, to avoid bias due to inconsistencies in the collection procedures employed in each country, i.e. we would not pool the data collected from different countries.
- Addition of relevant types of additive noise and convolutional channel effects to measure the impact that more realistic collection conditions would impose on the speaker recognition algorithms. Incorporating such realistic degradations could make subsequent experiments more relevant to real military applications.

Finally, we note that the N4 data set is well-suited for measuring the performance of speech recognition systems on tactical, non-native speech.

6. REFERENCES

[1] Benarousse L. et. al. “The NATO native and non-native (N4) speech corpus.” *Proceedings of Workshop on Multilingual Speech and Language Processing*, Aalborg, Denmark, 2001.

[2] Bimbot, F., Magrin-Chagnolleau I., Mathan, L. “Second-order statistical measure for text-independent speaker identification.” *Speech Communication*, Vol. 17, pp. 177-192, 1995.

[3] Ricart R., Cupples J., Fenstermacher, L.. “Speaker recognition in tactical communications.” Proceedings of *the International Conference on Acoustics Speech and Signal Processing*, Adelaide, Australia, 1994. Vol. 1, pp 329-332.

[4] Gersho, A. and Gray, R. M. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992, pp. 362-367.

[5] Unpublished internal AFRL Report, 2000.

[6] Reynolds, D. A. “Comparison of Background Normalization Methods for Text-Independent Speaker Verification.” *Proceedings of the European Conference on Speech Communication and Technology*, Rhodes, Greece, 1997. Vol. 2, pp. 963-966.

[7] Reynolds, D.A., Dunn, R.B., McLaughlin, J.J. “The Lincoln speaker recognition system: NIST EVAL2000.” *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000. Vol. 2, pp. 470-473.

[8] Reynolds, D. A. “Speaker identification and verification using Gaussian mixture speaker models.” *Speech Communication*. Vol. 17, pp. 91-108, 1995.

Phonetic Refraction for Speaker Recognition

Mary A. Kohler[1], Walter D. Andrews[1], Joseph P. Campbell[2], Jaime Hernández-Cordero[1]

[1]Department of Defense
[2]M.I.T. Lincoln Laboratory
m.a.kohler@ieee.org, waltandrews@ieee.org, j.campbell@ieee.org, jhernandez@ieee.org

Abstract

This paper describes a newly realized high-performance speaker recognition system and examines methods for its improvement. Innovative experiments early this year showed that phone strings are exceptional features for speaker recognition. The original system produced equal error rates less than 11.5% on Switchboard-I audio files. Subsequent research indicates that the equal error rate can be nearly halved by improving the feature extraction and score fusion methods. Pre-processing the speech files to remove cross-talk, improved techniques for combining scores, and gender-specific phone models each reduce the error rates significantly.

1 Introduction

Pronunciation is an elemental factor for human recognition of speakers. Converting the process by which humans recognize speakers to repeatable machine techniques is a challenging task that has not been successfully attempted, until now. By capturing phone sequences and using them to examine the acoustic phonetic details of different speakers, we can detect and exploit differences in pronunciation.

We develop a speaker-recognition system based only on phonetic sequences instead of the traditional acoustic feature vectors. Although the phones are generated based on the acoustic feature vectors, the recognition is performed strictly from the phonetic sequences created by the phone recognizer(s).

Our phonetic speaker recognition approach relies on phonetic recognizers in several languages to capture phone sequences, which are then used for modeling and recognizing speakers. By processing the speech files with phone recognizers of different languages, we produce refracted phonetic sequences that provide complementary information. Combining phone sequences from several languages not only provides improved performance and robustness, but also provides a degree of language independence similar to that of acoustic approaches.

2 NIST Extended Data Task

All the experiments described in this paper use data from the NIST 2001 Speaker Recognition Evaluation Extended Data Task. NIST's purpose in creating this task was to promote the exploration and development of new approaches to the speaker recognition challenge, such as the idiolectal characteristics reported in [3] that require larger amounts of training data than provided in previous evaluations.

For the 2001 evaluation, the entire Switchboard-I corpus was prepared for the Extended Data Task. Along with the audio data, NIST provided both Dragon System's automatic speech recognition transcriptions, and manual transcripts from the Institute for Signal and Information Processing. Both sets of transcripts were available for the entire corpus. All forms of data were permitted for training speaker models either alone or in combination.

The speaker model training data consisted of one, two, four, eight, and sixteen conversations. NIST employed a jackknife approach to rotate through the training and testing conversations to insure an adequate number of tests. Table I provides a breakdown, based on the number of training conversations, of the NIST Extended Data Task.

Table I: NIST Extended Data Description

Number of Training Conversations	Number of Unique Speakers	Number of Test Conversations
1	483	16429
2	442	15363
4	385	13777
8	273	10377
16	57	2696
Total	483	58642

For testing, the same options were available as in training. The recognition feature could be computed from either the acoustic data, the transcriptions or a combination of both. The number of test conversations for each set of training conversations is provided in Table I. The test set contains matched handset and mismatched handset conditions as well as a few cross-gender trials.

NIST provided data in two-channel sphere-formatted audio files. Analysis of the individual conversation sides revealed a considerable amount of cross-talk, which could potentially inhibit successful speaker recognition. Unlike other NIST data, the Switchboard-I files were not processed to remove echo. In [7] we processed the Switchboard-I data through NIST's echo-canceling software prior to

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

speaker recognition. All experiments in this paper use the echo-cancelled Switchboard-I files. This paper includes experiments for removing cross-talk in the phone sequences to potentially improve speaker recognition performance.

3 Algorithm Description

Phonetic speaker recognition is performed in four steps. First, a phone recognizer, in the given language, processes the test speech utterance to produce phone sequences. Then a test speaker model is generated using phone n-gram (n-phone) frequency counts. Next, the test speaker model is compared to the hypothesized speaker models and the Universal Background Phone Model (UBPM). Finally, the scores from the hypothesized speaker models and the UBPM are combined to form a single recognition score.

The single-language system is generalized to accommodate multiple-languages by incorporating phone recognizers trained on several languages resulting in a matrix of hypothesized speaker models. The system described in this paper used *M* speakers, *P* phone recognizers and *P* UBPMs, one UBPM corresponding to each phone recognizer. Figure 1 shows this multi-language phonetic speaker-recognition system. The following sections provide more details for the modeling and recognition process.

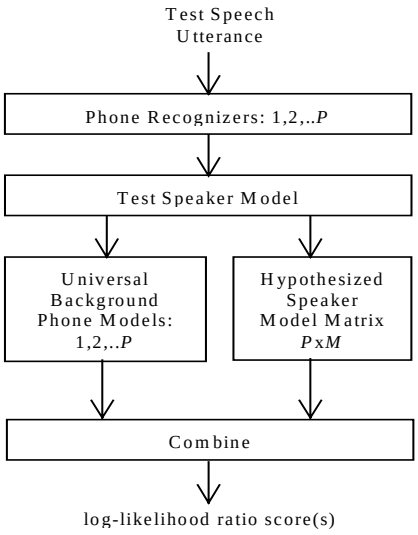


Figure 1. Multilanguage Phonetic Speaker-Recognition system

3.1 Phone Recognition

The phone recognition process takes advantage of a phone recognition algorithm that Zissman created for Parallel Phone Recognition with Language Modeling (PPRLM) [4]. We chose this recognizer since it was created solely for phone recognition with no language model constraints. This algorithm calculates twelve cepstral $(c_1 \cdots c_{12})$ and thirteen delta-cepstral $(c'_0 \cdots c'_{12})$ features on 20 ms speech frames with 10 ms updates. The cepstra and delta-cepstra are sent as two independent streams to fully connected, three-state, null-grammar HMMs.

The HMMs were trained on phonetically marked speech from the Oregon Graduate Institute (OGI) multi-language corpus in six languages: English (EG), German (GE), Hindi (HI), Japanese (JA), Mandarin (MA), and Spanish (SP). The corpus was hand-marked by native speakers in each language using OGI symbols for two of the languages and Worldbet symbols for the remainder. The number of phonetic symbols differs for each language from 27 for Japanese to 51 for Hindi, and includes one symbol to represent silence. More information on the corpus and phone symbols can be found in [8] and [9].

The phone recognizer employs a Viterbi HMM decoder implemented with a modified version of the HMM Toolkit. The output probability densities for each observation stream (cepstra and delta-cepstra) in each state are modeled as six univariate Gaussian densities. The output from the HMM recognizer for each language provides four estimates: the symbol for the recognized phone, its start time, its stop time, and its log-likelihood score. For this paper we only used the recognized phone, although future plans include exploiting the other estimates.

3.1.1 Gender-specific Phone Recognition

Zissman also created gender-specific phone models in five languages (EG, GE, JA, MA, SP) using the OGI multi-language phone-marked speech corpus. The phone models are identical in format to those described previously, but the training speech was constrained by gender. We conducted some preliminary experiments using gender-specific phone models to create gender-dependent phone sequences for speaker recognition.

3.2 Cross-talk Removal

As described previously, the original Switchboard-I audio files contained excessive cross-talk. Since this interference was potentially deleterious to speaker recognition performance, we elected to process the audio files with software created by MIT Lincoln Laboratory. Their *xtalk* tool performs energy based cross talk and silence detection, producing separate files containing speech marks and speech.

We separated the conversation sides from the raw stereo Switchboard-I files prior to *xtalk* processing. In a time-saving effort, we chose not to run the time-intensive phone-recognition software on the *xtalk*-processed speech. Instead, we converted the existing phone files (created from echo-cancelled Switchboard-I files) using the two-channel speech activity detection

(SAD) marks to determine whether the phone should exist. Figure 2 shows this procedure in more detail.

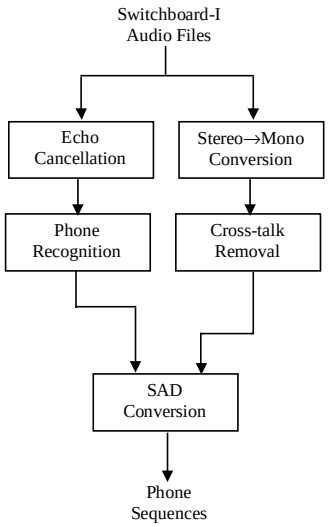


Figure 2. Cross-talk Elimination Process

We experimented with several thresholds to determine when a phone should be included. We found that the best speaker recognition performance was achieved by including all phones occupying any portion of a valid speech segment. The converted phone files were processed by the back-end for speaker recognition as described below.

3.3 Utterance Delineation

Previous work [6], [7], showed that processing the phone files to include start and stop tags around speech phrases improved speaker recognition performance. The previous algorithm inserted start and stop labels between phrases based on pairs of silence phone labels, i.e., all phones between two silence phone labels were considered an utterance. For example, if the recognized phone sequence was

… sil S oU m i: D & m Λ n i: sil …

the utterance-tagged speech became

…<start> S oU m i: D & m Λ n i: <end> <start>…

regardless of the length of the silence phones.

In this paper we analyzed the distribution of silence phone durations and experimented with more sophisticated methods to determine where to place utterance breaks, as described later.

3.4 Hypothesized Speaker Model

As noted in section 2, a jackknife scheme determined the amount of training and testing data for the extended training task. NIST provided a control file listing hypothesized and test speakers, along with a training and testing conversation list [5]. The list provided training information for one, two, four, eight, and sixteen conversations. As a result, a particular hypothesized speaker will have multiple models for a given test set.

Speaker dependent language models, *H*, are generated using a simple n-phone frequency count for each language and consist of all the unique n-phones with the corresponding frequency counts for a given speaker. Unlike the state-of-the-art GMM-UBM systems, the speaker models are not adapted from the UBPM.

3.5 Universal Background Phone Model

The UBPM, *U*, is generated using files determined from the NIST control file (specified in [5]), which provides a list of hypothesized and test speakers for exclusion from the UBPM. All of the conversations for the remaining speakers were used to build the UBPM using n-phone frequency counts. For this paper, each of the six phoneme recognizers has a corresponding independent UBPM.

3.6 Test Speaker Model

A test set is specified in the NIST control file for all hypothesized speaker models. The test set contains true speaker trials, impostor trials, matched handset, mismatched handset, and a few cross-gender trials. Once the speech utterance to be tested is processed by the phone recognizer(s), a test speaker model, *T*, is generated using n-phone frequency counts. Doddington, [3] improved performance by ignoring infrequent word n-grams, i.e. ignoring n-grams occurring less than $c_{\min}$ times. This is also the case with the phonetic approach.

3.7 Scoring

Producing a speaker recognition score for a speech file requires not only the calculation of a likelihood score for each of the *P* phone recognizers, but also the combination of each of the *P* scores into a single score for comparison with other hypothesized speaker model scores. The following sections describe how the individual model score is calculated and fusion of the *P* model scores.

3.7.1 Single-language Scoring

For a single-language phonetic speaker-recognition system, the scores from the hypothesized speaker models and the UBPM are combined to form the recognition score η_i using a conventional log-likelihood ratio given by

$$\eta_i = \frac{\sum_n \left(w(n) \left[S_i(n) - B(n) \right] \right)}{\sum_n w(n)}$$

where n is an n-phone type corresponding to the test speaker model, T, and the sums run over all of the n-phone types in the test segment, T. S_i represents the log-likelihood score from the i^{th} hypothesized speaker model, H_i, and B is the log-likelihood score from the UBPM, U, for the n-phone type, n. The log-likelihood scores S_i and B are defined by

$$S_i(n) = \log\left[\frac{H_i(n)}{N_{H_i}} \right] \quad and \quad B(n) = \log\left[\frac{U(n)}{N_U} \right],$$

where N_{H_i} and N_u represent the total number of unique n-phone types in the i^{th} hypothesized speaker model and UBPM, respectively. $H_i(n)$ and $U(n)$ represent the number of occurrences of a particular n-phone type, n, in the hypothesized speaker model and UBPM, respectively.

The weighting function $w(n)$ is based on the n-phone token count, $c(n)$, and the discounting factor, d. The n-phone token count, $c(n)$, corresponds to the number of occurrences of a particular n-phone type n in the test speaker model, T. The weighting function, which could be made language dependent, is given by

$$w(n) = c(n)^{1-d}.$$

The discounting factor, d, has permissible values between 0 and 1. When $d = 1$ a complete discounting occurs, resulting in $w(n) = 1$. This gives all n-phone types the same weight regardless of the number of occurrences in the test speaker model, T. When $d = 0$, all n-phone types are weighted by their corresponding token count in the test speaker model, T.

3.7.2 Multiple-language Scoring

In [7], the scores from each of the single-language phonetic speaker-recognition systems were fused by a simple linear combination. Subsequent experiments revealed that more sophisticated techniques for combining the individual language scores improved phonetic speaker recognition performance. We used the LNKnet tool developed by MIT's Lincoln Laboratory to experiment with several different classification techniques using vectors of the individual language scores as features.

The Gaussian mixture classifier showed the most promising results using either expectation-maximization binary split or K-means for clustering. Both clustering algorithms used eight mixtures, a grand/class full covariance matrix, and tied mixture components.

4 Results

The preceding sections described several experiments intended to improve the speaker recognition performance described in [7]. The results from these experiments are presented below. All detection estimation tradeoff (DET) curves shown are for systems trained on eight conversations with triphone models $(n = 3)$, complete discounting $(d = 1)$, and ignoring n-phones that occur less than 1,000 times, $(c_{\min} = 1000)$. Unless noted otherwise, individual scores from six languages $(P = 6)$ were linearly combined with equal weights to calculate the final phonetic speaker recognition score.

4.1 Utterance Delineation

Analysis of the distribution of silence phone lengths using duration information calculated from the phone recognizer output showed that most of the silences were of short duration (less than 200 ms). The existing approach, which used silences of any duration to mark an utterance, seemed inefficient. Since the goal of utterance delineation was to accurately separate speech phrases, the best approach seemed to use only long silences as utterance separators.

We experimented with several thresholds for minimum silence duration to denote utterances, from 300 ms to 1.2 s. When we compared the speaker recognition performance of the more discriminatory techniques with the simple, naïve approach, we found that speaker recognition performance did not improve with the more complex methods.

4.2 Two-channel Speech Activity Detection

We performed two speech activity detection experiments to determine the optimal method for removing cross-talk. In one experiment, the two-channel SAD processing was performed on the original Switchboard-I files. In the other experiment, two-channel SAD processing was performed on the echo-cancelled Switchboard-I files. For both experiments, we converted the existing phone sequences created from echo-cancelled Switchboard-I files, as shown in Figure 2.

Figure 3 and Figure 4 show DET curves of speaker recognition performance for phones processed with two-channel SAD experiments and for phones processed only on echo-cancelled files. Figure 3 shows performance using only an English phone recognizer, and Figure 4 shows performance using six phone recognizers. On both figures, the solid line marks performance for the experiment in which the original Switchboard-I files were used to create two-channel

SAD marks. The dotted line shows performance for the experiment in which the echo-cancelled Switchboard-I files were used to create two-channel SAD marks, and the dash-dot line indicates performance without two-channel SAD processing. The boxes on each plot demarcate the 90% confidence interval around the equal error rate (EER) based on the number of target and non-target trials.

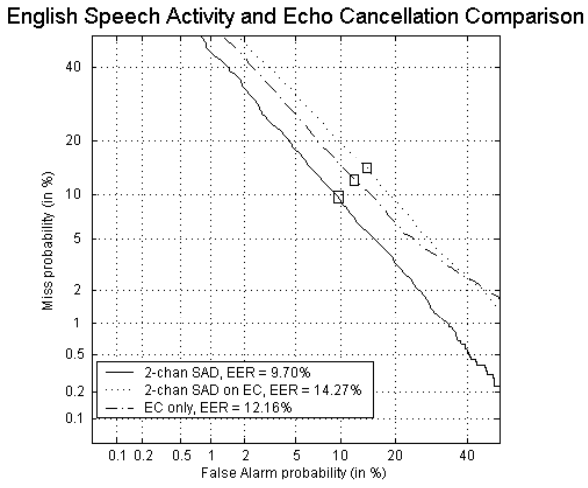


Figure 3. Comparison of Three Audio Processing Techniques, English

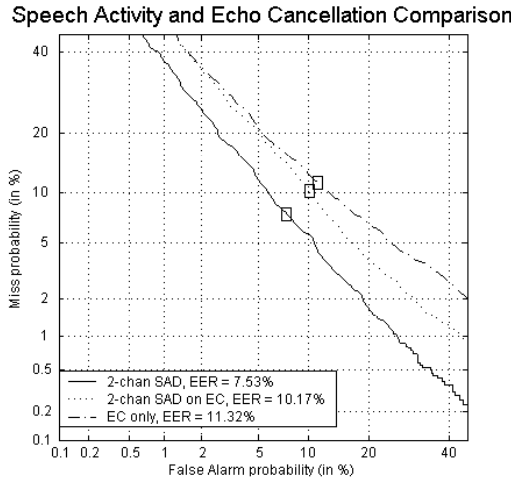


Figure 4. Comparison of Three Audio Processing Techniques, Six Language Fusion

The data in Figure 3 and Figure 4 demonstrates that cross-talk removal improves the EER for speaker recognition performance by nearly 2.5% in the English-only case and more than 3.5% in the combined language case. It also shows that using the unprocessed speech files to determine the regions of speech is superior to using echo-cancelled speech to determine these regions. We plan additional audio processing experiments to discover future recognition improvements. The following experiments use this cross-talk removal process to modify the phone sequences.

4.3 Language Fusion

NIST provided six splits of the speech data each containing unique speakers. We used vectors of the *P* language scores from two of these splits to train the classifiers through LNKnet and two splits to test the classifiers. Figure 5 shows a comparison of the linearly combined phone scores with the Gaussian mixture classifiers. The dotted line and solid line show performance for K-means clustering and expectation-maximization clustering of the Gaussian mixture models, respectively. The dash-dot line shows performance for a linear combination of the *P* language scores, and the dashed line shows performance using only an English phone recognizer. The boxes on each plot demarcate the 90% confidence interval around the equal error rate based on the number of target and non-target trials.

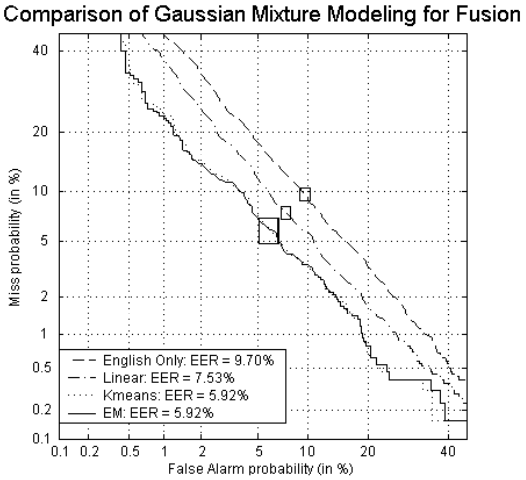


Figure 5. Comparison of Two Gaussian Mixture Models for Fusing Language Scores

As Figure 5 shows, both Gaussian mixture classification techniques are superior to linear fusion, but are nearly identical in performance to each other. Additional experiments to determine improved methods for combining the individual language scores are planned for the future.

4.4 Gender-Specific Phone Modeling

NIST provided a table with the Switchboard-I files containing information about each of the audio files including the speaker's gender. We processed the audio files with the five gender-specific phone recognizers described previously. Figure 6 contains a comparison for speaker recognition performance using

only an English phone recognizer with and without gender-dependent phone models. The solid line shows performance for phone recognition using separate phone models for males and females. The dotted line shows speaker recognition performance using the same phone model regardless of gender. The boxes on each plot demarcate the 90% confidence interval around the equal error rate based on the number of target and non-target trials.

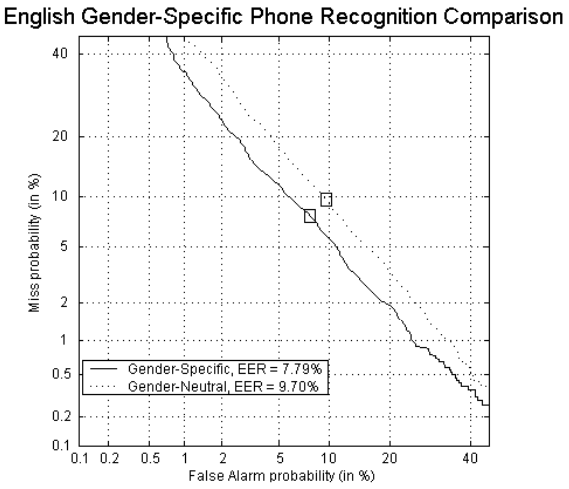


Figure 6. Comparison of Gender-Dependent and Gender-Independent Phone Recognition for English

As shown by the data in Figure 6 speaker recognition performance can be improved significantly in the one-language case using phone models to match the speaker's gender. Further experimental results using phones from five language will be reported in future publications.

5 Conclusions

This paper described an innovative technique for speaker recognition using phonetic sequences to capture a speaker's pronunciation. Four improvements to the basic model were described, most of them exploiting front-end signal processing.

Sophisticated methods for separating speech phrases did not outperform the simple, original method. This was an unexpected result, but its simplicity saves computation.

Elimination of cross-talk provides a significant improvement, especially when the unprocessed speech files are used to determine the areas containing speech. The equal error rate decreased by nearly 4% when the phone strings taken from the echo-cancelled data were post-processed to remove cross-talk.

Gaussian mixture classification for fusing individual language scores is an improvement over a linear combination of the scores. K-means and expectation-maximization binary split clustering perform essentially identically. They each provide over 1.5% improvement in equal error rate.

Gender-specific phone models are superior to using a combined phone model for English phone strings. The equal error rate for the phonetic speaker recognition system decreases by almost 2% when the gender of the speaker is matched to the gender of the phone model.

Additional experiments to improve performance are underway. Improved front-end and back-end processing as well as investigation of other feature sets will be reported as we collect results.

6 Acknowledgements

The authors thank George Doddington and John Godfrey for their helpful discussions.

7 References

[1] Reynolds, D., T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19-41.

[2] Weber, F., B. Peskin, et al., "Speaker Recognition on Single- and MultiSpeaker Data," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 75-92.

[3] Doddington, G., "*Some Experiments on Idiolectal Differences Among Speakers*," http://www.nist.gov/speech/tests/spk/2001/doc/, January 2001.

[4] Zissman, M., "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Trans. On SAP*, vol. 4, Issue 1, January 1996.

[5] Przybocki, M., and A. Martin, "*The NIST Year 2001 Speaker Recognition Evaluation Plan*," http://www.nist.gov/speech/tests/spk/2001/doc/, March 1, 2001.

[6] Andrews, W., M. Kohler, and J. Campbell, "Acoustic, Idiolectal, and Phonetic Speaker Recognition," To appear, *Proceedings of 2001: A Speaker Odyssey, The Speaker Recognition Workshop*, Chania, Crete, Greece, June 18-22, 2001.

[7] Andrews, W., M. Kohler, and J. Campbell, "Phonetic Speaker Recognition," To appear, *Proceedings of Eurospeech 2001*, Aalborg, Denmark, September 3-7, 2001.

[8] Lander, T. and S. Metzler, "*The CSLU Labeling Guide*," CSLU Oregon, February 1994.

[9] Hieronymus, J., "ASCII Phonetic Symbols for the World's Languages: Worldbet," *Journal of the International Phonetic Association*, 1993.

Methods and Models for Quantitative Assessment of Speech Intelligibility in Cross-Language Communication

Sander J. van Wijngaarden, Herman J.M. Steeneken and Tammo Houtgast

Department of Perception
TNO Human Factors, PO Box 23, 3769 ZG Soesterberg, The Netherlands
{vanWijngaarden, Steeneken, Houtgast}@tm.tno.nl

Abstract

To deal with the effects of nonnative speech communication on speech intelligibility, one must know the magnitude of these effects. To measure this magnitude, suitable test methods must be available. Many of the methods used in cross-language speech communication research are not very suitable for this, since these methods are designed to investigate specific effects regarding speech perception and production, rather than quantifying overall intelligibility. In this paper, a simple model of cross-language speech intelligibility is shown that helps in selecting experimental methods to assess speech intelligibility. Based on this model, and practical observations regarding assessment of cross-language speech intelligibility, a multi-lingual version of the Speech Reception Threshold method was implemented as a suitable method for the quantification of cross-language speech intelligibility. The performance of this method is illustrated by means of experimental results.

1. Introduction

Most reported experiments concerning nonnative speech intelligibility have been designed to obtain a better insight into the details of the speech perception and production process. Researchers in the field of second-language speech production and perception usually aim to test very specific hypotheses. Which experimental method is the most efficient depends on the tested hypothesis.

Apart from research on the basics of human speech communication an increasing need is felt for a more applied approach, aiming at the overall effect on speech intelligibility. Cross-language speech communication, in which one or more parties engaged in a conversation depend on second-language skills, is an increasingly common phenomenon. The efficiency of cross-language speech communication is quite often experienced to be lower than 'fully native' communication. For many of those situations, it would be helpful to be able to assess the magnitude of the effect on speech intelligibility. Applications that could benefit from such knowledge would be, for example, the design of public address and communications systems, and prediction models in room acoustics. By knowing the extent to which speech intelligibility is reduced, better design criteria can be established.

Wanting to know the *extent* to which speech intelligibility is influenced means that quantitative methods for measuring speech intelligibility are needed. This is different from the hypothesis-driven methodology preferred for investigating the principles of nonnative speech communication; instead of looking for *reasons,* we are quantifying the *consequences.*

To illustrate this approach, consider the following situation. Suppose that an auditorium in a Dutch school is equipped with an air-conditioning system, which produces a known level of background noise. In 'normal' (native) situations, the intelligibility of the public address system in the auditorium is generally acceptable, despite the background noise. What if a native English talker addresses the Dutch students (in English), who have an average experience with the English language of 2 years? What if the average experience of the students is 5 years, or what if the native language of the talker is German? What reduction of the background noise level is necessary to obtain a certain minimum speech intelligibility?

When using suitable methods, it is possible to answer all these questions, if populations of talkers and listeners are properly defined. Not all of the *reasons* behind the differences in intelligibility have to be known. These reasons may be very complex, involving better analysis of the speech signal into phonetic units, larger vocabulary, better understanding of the grammar, etc. Regardless of the reasons, the effects are interesting enough in their own right.

In this paper, we will present a simplified model of nonnative speech communication. The aim of this model is to serve as a tool, which helps in choosing the proper methods to quantify the effects on intelligibility. Based on this model, we will describe a multi-lingual speech intelligibility evaluation method that is suitable for application to cross-language speech communication

2. Model of cross-language speech communication

2.1. Types of cross-language speech communication

Describing a specific cross-language conversation unambiguously takes a little consideration. As the number of people engaged in a conversation increases, the complexity of a proper description of the situation increases accordingly.

All situations can be broken down into variants of straightforward two-way communication, in which case only one person is talking, and only one other person is listening. This involves influences from up to three languages: the native language of the talker, the native language of the listener, and the language that is currently being spoken. The relations between these three languages will partly determine the speech communication process. Comparative studies of the involved languages could theoretically shed light on this; analyses of the existence of phonetic contrasts and inspection of the (sound-based) lexicon of a specific language could help understand its relation with other languages, provided this

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

same information is also known for these other languages. Rather than trying to find a general model for language-related influences on cross-language communication, we will treat each combination of languages as a unique case.
It has become convention to denote native talkers and listeners as 'L1', and nonnative (second-language) talkers and listeners as 'L2'. Based on this notation, one could (for example) indicate that a native listener is listening to a non-native talker by writing 'L2>L1'. This notation works if the number of languages involved is no more than two. The situation 'L2>L2' could mean that a Dutch listener is speaking English to a German listener; it could also mean that a Dutch listener is speaking English to another Dutch listener. The difference may be important, since the common native language between talker and listener may influence their use of the second language (in our example English).
To avoid confusion, we will use the following notation throughout his paper:

Dutch > (English) > German

meaning that a Dutch talker is talking English to a German listener. We will generally abbreviate this to D>(E)>G.

2.2. Defining populations of talkers and listeners

Considering nonnative speech intelligibility separately for each individual that comes our way would become a very laborious process. By defining meaningful populations of talkers and listeners, we can collect more generally applicable quantitative results. First, we decide what populations we need to have quantitative data on; then we recruit subjects from these populations, and carry out experiments. Experiments may involve subjects selected from one single population, or may use talkers from one population and listeners from another.
In order to define a population, one should be able to describe it in terms of the determining factors for nonnative speech intelligibility. The description of the population starts with the native language of the subjects; preferably, details concerning regional accents (if any) should also be known.
A very important factor is the average experience of subjects within the population with the target (second) language (eg. ([1,2]). Age of acquisition of the second language is also of great importance. (eg. [3,4,5]).
Second-language experience and age of acquisition combine into second language *proficiency*, a term we will use rather loosely to indicate the underlying dimension explaining differences in nonnative speech intelligibility. Despite the fact that second-language proficiency comprises many different abilities (related to phonetic discrimination, vocabulary, grammar, etc.), subjects are able to rate their own proficiency with a sometimes impressive accuracy [6].
Possible other factors to consider could be more general descriptors of the population, such as age and gender. It seems fair to consider the influence of these variables on cross-language communication higher-order effects, but it is only prudent to keep variables like these in mind as well when selecting subjects for experiments.
Even when the populations of talkers and listeners are fully defined, the resulting speech intelligibility may still vary according to numerous other variables, most of which also apply to fully *native* communication, such as speaking rate and speaking style. These variables are not really related to the characteristics of the talkers and listeners, but rather to their *mode* of communication. One aspect related to this is worth mentioning. For nonnative talkers, the distinction between *read* speech and *spontaneous* speech is potentially of far greater importance than for native talkers. Nonnative talkers are likely to limit their effective vocabulary to easier and more familiar words when speaking spontaneously, while they are more likely to produce pronunciation errors when asked to read a certain text aloud. In the latter case, they are not only likely to mispronounce unfamiliar words, but a poor understanding of context may also lead to an impaired intonation of sentences.

2.3. Conditions for speech communication

Native as well as nonnative speech can be affected by adverse conditions, such as background babble, ambient noise, bandwidth limiting, or reverberation. However, the degrading influence on cross-language speech communication tends to be greater [5,7,8,9,10].
Measuring speech intelligibility under clear, undegraded, conditions is often not very effective. The effects of nonnativeness on intelligibility may be relatively small, whereas problems in practice are expected when degrading circumstances *are* present. By conducting experiments under conditions that represent a controlled degree of speech signal degradation, the effect of this degradation on cross-language speech communication may be assessed systematically.
Perhaps the easiest way to reduce speech intelligibility in a controlled manner, is by adding stationary noise with a known spectrum. For fully native speech communication, intelligibility in this case is a relatively stable and well-known function of the speech-to-noise ratio. For nonnative speech communication similar relations are found [10,11], which clearly show that noise is capable of affecting cross-language communication more profoundly than native speech communication.

2.4. Levels of analysis

Our approach towards the assessment of nonnative speech intelligibility needs a model that describes cross-language speech communication in such a way, that the proper characteristics to quantify intelligibility can be chosen.
In practice, this means that a description is needed of the determining factors for speech intelligibility (which we will call intelligibility cues), and an indication of where to find these. More specifically, we need to find out about intelligibility cues that are especially important when considering *cross-language* speech communication.
Speech intelligibility can be studied at various levels of analysis; the most basic analysis would involve studying the speech signal on an allophone-by-allophone basis. Perhaps the highest thinkable level would be to consider an entire story, where the amount of relevant information in the story that was transferred could be studied.
There are reasons to assume that the level of individual words takes an important position in the process of learning a second language [12]; it seems likely that one initially learns a second language mainly by collecting a sound-based representation of its lexicon. For this reason, and because of practical considerations, we will distinguish three levels of analysis: speech units smaller than words (allophones), words, and speech units larger than words (sentences).

Besides the level of analysis, intelligibility cues can also be separated depending on whether they can be found in the speech signal ('acoustic' cues) or somewhere else. As an example of the difference: the intelligibility of sentences (as compared to the intelligibility of the individual words of which they consist) is enhanced by means of intonation. Intonation (or more generally, prosody) is present in the speech signal, and can therefore be called an 'acoustic' intelligibility-enhancing factor. The semantic and syntactic redundancy contained in a sentence also increases its intelligibility relative to the individual words of which it consists. However, these factors can not be traced back to the speech signal; they improve intelligibility by aiding the listener in his cognitive processing of the message.
Table 1 illustrates the distinction between acoustic and non-acoustic intelligibility cues at the three defined levels of analysis.

Table I. Levels of analysis in nonnative speech communication

Level of analysis	Examples of affected intelligibility cues	
	Acoustic	Non-acoustic
Supra-word level (sentence level)	Prosody	Syntactic constraints Semantic constraints
Word level	Lexical dissimilarity	Word familiarity
Sub-word level (allophone level)	Phoneme inventory	

This distinction between acoustic and non-acoustic factors is not helpful at the sub-word level. For the non-acoustic factors at this level (such as the individual phoneme space representation that a listener uses to categorize L2 allophones) can hardly be tested without involving acoustic allophone realizations.
Table 1 can be used to decide which *characteristic* of cross-language speech intelligibility is the most appropriate in a specific case, for instance phoneme recognition versus sentence intelligibility. Only *after* deciding which is the most appropriate characteristic can we design a proper experiment.
For example, one may wish to quantify the intelligibility of a group of (nonnative) German actors, playing before an audience of native English listeners, in the English language (G>(E)>E). The non-acoustic intelligibility cues do not require special attention in this case, since only the talkers are nonnative, and their vocabulary and sentence construction are 'programmed' by the play they are acting out. Hence, all deviations from fully native communications can be found in the speech signal. At the very least, one may expect that the actors' allophone realizations will deviate from native English speech. A phoneme-based intelligibility test will be a suitable choice to quantify this effect. However, this may not be the *most* suitable intelligibility test. Unless the actors are thoroughly trained by a native English director or language coach, their intonation will also deviate from the authentic English patterns. In that case, a (sentence-based) intelligibility test that is sensitive to differences in prosody is a better choice.
As another example, consider the reverse situation (the actors are now English and the audience is German; E>(E)>G). Since the German audience is now the only nonnative factor, the speech signal is not at all affected. Still, the resulting speech intelligibility may be reduced considerably; partly because the nonnative listeners are not as good at identifying individual speech sounds, but also for reasons related to vocabulary and the less effective use of word context [11]. In this case, the average L2 linguistic development of the German audience is an important variable. Besides a speech intelligibility test using sentences (to include the effects of word context), it may be useful to include a separate test to quantify vocabulary and context-effects separately.

3. Speech intelligibility assessment methods

3.1. Practical considerations

A pragmatic approach toward measuring nonnative speech intelligibility is simply to adopt one of many proven experimental methods designed for native speech. Inevitably, some modifications to these proven methods will be necessary, if only for practical reasons.
Several intelligibility test methods are based on one-syllable nonsense words. These tests are generally quite efficient at measuring speech intelligibility phoneme level. Subjects participating in such tests must somehow communicate perceived nonsense syllables in response to the auditory stimuli. With L2 listeners, typing these responses should be ruled out as an option. Differences in orthographic representations of sounds between L1 and L2 will confuse the subject. Even highly proficient subjects, who are aware of differences in orthography between L1 and L2, are likely to produce errors, especially when working under time pressure. Collecting multiple-choice responses will partly solve this problem, especially if no 'confusing' alternatives are presented. In any case, proper subject instruction with regard to this issue is vital.
Some additional complications surrounding experiments with non-natives have to do with the recruiting of subjects. The definition of the population from which to draw subjects is much narrower than usual in speech intelligibility testing. Accordingly, subjects will be harder to find. Experimental methods can be designed or adapted to help cope with this issue. Methods that require special sound-insulated rooms or heavy equipment require subjects to travel to a certain location. By adapting these methods so that they can be implemented in a portable device (such as a notebook computer) hard-to-reach subjects (unwilling to travel in order to take part in a test) can be tested at remote locations.
The available time per subject may also be shortened. When tests run over longer periods of time, a smaller percentage of the population of potential subjects will be willing to participate. By shortening the duration of the experiment (by making tests more efficient, or by spreading the load over a slightly larger number of subjects) the number of available subjects may be increased.

3.2. Types of speech stimuli

Various types of speech stimuli are used in speech intelligibility tests. Generally, the length of each single stimulus determines which level of analysis (table I) is addressed by the test method.
The most fitting speech stimuli corresponding to the different levels indicated in table I would appear to be sentences, words and phonemes. However, individual phonemes are hard to test without the context of a word or syllable; hence the

frequent use of nonsense syllables that was mentioned in the previous section. The individual recognition of phonemes is also difficult to test using *meaningful* words, since the word context will be of some influence on the probability of correct recognition.
Higher-then-word level effects are expected for most thinkable cross-language conversations. In principle, sentence intelligibility tests also include effects at lower (word and phoneme) levels, since all sentences are constructed from these smaller units of speech. If only one type of speech stimuli can be chosen, it makes sense to choose sentences. On the other hand, it should be noted that (nonsense) word tests will be more sensitive to effects at lower levels of analysis.
When comparing native and nonnative talkers, specific choices must be made before recording any speech stimuli. Speaking rate and speaking style are likely to vary between native and nonnative talkers. Nonnative talkers usually tend to (consciously or unconsciously) compensate for the effects of their accent on intelligibility by adjusting their speaking rate or speaking style [6]. This is a legitimate effect, which can also be observed in cross-language conversations in practice – it is in some ways similar to the Lombard-effect, which lets talkers automatically increase their vocal effort in the presence of background noise. One may choose to include this effect in the test, or force native and nonnative talkers into similar speaking styles (by giving suitable instructions, monitoring recordings, and pacing their speaking rate).

3.3. Multi-lingual test methods

One step further than nonnative speech intelligibility testing is multi-lingual intelligibility testing. Multi-lingual tests can involve either native or nonnative subjects, but must also be implemented in multiple languages. Obtaining equivalent implementations of the same test in various languages poses an additional difficulty. True equivalence across languages is hard to reach.
Whatever speech stimuli are used, these stimuli must somehow be matched across languages. When working with phoneme tests, the tested phonemes could be balanced to represent the mean frequency of occurrence in the corresponding language. Despite the fact that different phoneme sets must be tested for each language, these are equivalent in the sense that they represent a 'natural' distribution of phonemes for each languages.
When the test stimuli are isolated words, then on top of phonetic balancing the frequency distribution of the test vocabulary (measured frequencies of occurrence in representative texts) should be controlled. Where available, the appropriate information could be taken from (multi-lingual) lexical databases.
When using sentences, the main things that should be matched across sentences are the complexity of the sentences, and the *domain* from which the sentences are taken. The source of the sentences largely determines the domain (newspaper, radio, everyday conversation, etc.), making this variable relatively easy to control. The complexity can be controlled by adopting certain constraints for the selection of sentences; at least the length (number of syllables) of the sentences, and the length of the individual words in the sentences, should match pre-defined criteria.
When sentences are properly selected, phonetic balancing becomes of lesser importance. Each sentence consists of a certain mix of phonemes; when each condition is tested with multiple sentences, there is a more or less implicit phonetic balancing for the domain from which the sentences are taken.
An additional complicating factor when designing multi-lingual tests is the fact that the relative importance of different levels of analysis (table I) may vary between languages. Phoneme identification may be more difficult in some languages than others, simply because the number of existing phonemes differs (eg. English vowels versus Spanish vowels). Contextual information that is available in one language, for instance by the use of case and word gender, may be absent in other languages.
A pragmatic approach to the design of multi-lingual test is to simply try out the implementations in different languages on native subjects. If the native scores are the same across languages, then it seems fair to assume that the method performs equivalently.

3.4. Multi-lingual Speech Reception Threshold method

An example of a multi-lingual implementation of an existing intelligibility test method is the multi-lingual Speech Reception Threshold (SRT) method. The SRT method is widely used as a diagnostic tool in the field of audiology [13], and has been proven useful to evaluate speech intelligibility of talkers, listeners, and communication systems.

3.4.1. Test procedure

The SRT 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 speech at a known speech-to-noise ratio. The masking noise spectrum is equal to the long-term average spectrum of the test sentences. After presentation of each sentence, the subject responds by orally repeating the sentence to an experimenter. The experimenter compares the response with the actual sentence. If every word in the responded sentence is 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 of a list of 13 sentences 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 over the last 10 sentences, the 50% sentence intelligibility threshold (SRT) is obtained.

3.4.2. Interpretation of SRT results

The score resulting from an SRT test ('the SRT' for the corresponding condition) is a speech-to-noise ratio (SNR); at this SNR, 50% of the sentences are repeated correctly by the listeners. At better (higher) SNRs, more than 50% will be intelligible, at more adverse (lower) SNRs, less than 50%. A lower SRT means better intelligibility: more noise can be allowed to reach 50% recognition of sentences.
The percentage of correctly recognized sentences is a (psychometric) function of SNR, often modeled as a cumulative normal distribution. The SRT is the adaptively estimated mean of this distribution, which is the best single parameter to characterize the whole curve. A logical second parameter to estimate would be the variance of the distribution, reflected by the slope of the psychometric curve. To estimate this slope (or even the full psychometric curve), one could use alternative testing paradigms using the same

SRT sentences. The description of such methods is beyond the scope of this paper.

3.4.3. Creating a multi-lingual version

The 'original' [13] Dutch SRT sentences describe common, everyday situations in simple wording. Based on these original sentences, the following constraints were defined for 'translation' of the sentence material:

- Sentence length 7-9 syllables
- No words longer than 3 syllables
- No more than one three-syllable word per sentence.
- Sentence content is of an everyday life nature
- Sentences of approximately equal redundancy (or predictability, perplexity) as the original sentences

3.4.4. Software implementation

A computer program was developed for maintaining multi-lingual databases of recorded SRT sentences and using these in intelligibility tests. This program also features a module for recording new material. In combination with a notebook computer and a high-quality sound card, a small, flexible setup is create which can be used to record and test talkers nonnative and listeners at any location that is sufficiently silent.

3.4.5. Speech recordings

Traditionally, talkers used in SRT tests for audiological purposes are trained professionals, speaking very clearly. The SRT scores obtained with these recordings are hard to reach for most ordinary talkers, especially under representative conditions.
Multi-lingual SRT talkers are not selected according to a strict regime, or following specific criteria. The talkers are simply verified not to exhibit any speaking disorders, and instructed to speak with a clear 'reading voice'. This makes it easier to recruit talkers, and quickly build up speech databases.
To prevent large differences in speaking rate, the speaking rate is paced by means of a 'progress bar'. Talkers have to pronounce each sentence within a 2.5-second timeframe, which is visually indicated on the computer screen.

3.4.6. Applications of the multi-lingual SRT

It should be noted that the application of the SRT method (and similar methods) to cross-language research is not new (eg. [5,14]). What is new about our current multi-lingual SRT implementation, is the effort to construct a coherent test in as many languages as possible. At the moment this paper was written, 'translations' of the sentences (text) were available in at least 8 different languages; a multi-speaker test speech database had been collected for at least 5 of these languages.
Sofar, the English, German and Dutch versions of the test were succesfully used to quantify cross-language speech intelligibility [6,11]. Apart from this, the English, French, German and Dutch version were used with solely native subjects (talkers and listeners) to measure the language dependency of voice coding systems [15].

4. Examples of experimental data

4.1. Nonnative listeners

For a population of Dutch university students, cross-language intelligibility-effects (in terms of SRT) were measured when listening to English and German [11]. Almost all Dutch university students have been taught English and German during secondary education, German at a slightly later age and for a shorter period than English. Also because of the more frequent use of English (university classes, textbooks, television and other media) the L2 proficiency tends to be much higher in English than in German. Figure 1 shows native and non-native SRT results related to this population.

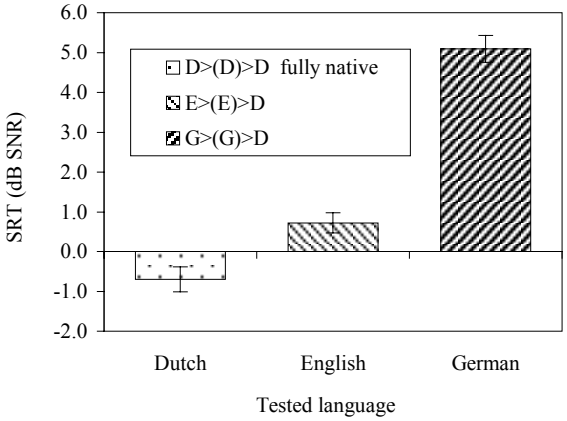


Figure 1. Mean SRT scores and standard errors measured for 9 Dutch university students when listening to three languages (3 talkers per language, N=27). (underlying data previously published, [11]).

The difference between the effects of listening to English and German is considerable; all differences in figure 1 are statistically significant. Despite the fact that the listeners were selected to be highly proficient in English, the effect of being nonnative listeners on the resulting intelligibility is clearly noticeable.

4.2. Nonnative talkers

Similar results as presented in figure 1 can be obtained for nonnative talkers. In that case, the population of listeners consists of 'average natives', and the talkers are recruited to match a certain desired profile.
Figure 2 shows results of in experiment aimed at measuring the effect of perceived foreign accent on intelligibility. For this experiment, 15 talkers were recruited who could all speak Dutch, differing in degree of foreign accent. These talkers were from 5 language backgrounds: Dutch, English, German, Polish and Chinese.
To measure the 'degree of perceived accent', a pairwise comparison experiment was conducted with native Dutch listeners. From this experiment, subjective foreign accent ratings were calculated. The relation between SRT scores and these ratings are shown in figure 2.

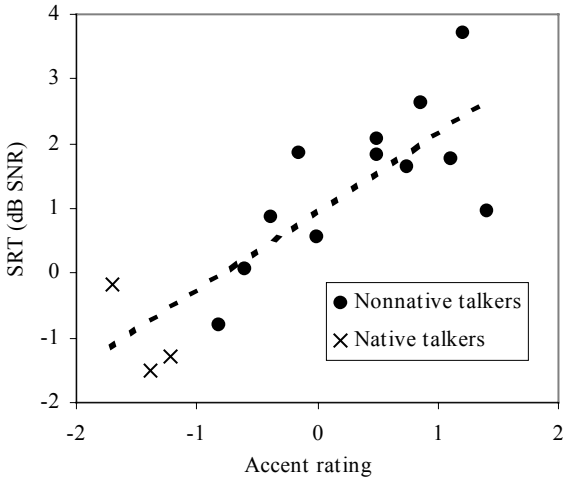


Figure 2. Relation between subjective accent ratings (N=39) and SRT scores (10 native listeners for each data point). R^2=0.68. Data previously presented [6].

Following from the correlation in figure 2, 68% of the variance in SRT scores could be explained by the perceived accent ratings.

4.3. Multi-lingual comparison

For some goals, multi-lingual speech intelligibility tests are useful even when no cross-language factors are directly involved. For example, to measure the language dependency of vocoders, one needs to test (native) speech intelligibility in a number of languages. The performance of a multi-lingual test should be closely matched across languages, otherwise the language dependency of the tested vocoders will be confounded with the language dependency of the test method [15]. The easiest way to verify if the results are sufficiently closely matched across languages, is by measuring the same (relatively undistorted) conditions in several languages. Results of such an experiment are shown in table II.

Table II. Mean native SRT scores and standard errors for four languages (3 talkers per language, 10 listeners). All speech was bandwidth limited (50-4000 Hz).

SRT	English	French	German	Dutch
mean	0.7 dB	1.0 dB	0.3 dB	0.4 dB
S.E. (*N*=10)	0.4 dB	0.8 dB	0.4 dB	0.5 dB

The results are closely matched (also note the magnitude of the effects shown in figures 1 and 2). None of the differences between languages are statistically significant.

5. Discussion and conclusion

The pragmatic model of cross-language speech communication presented in this paper was used to select the multi-lingual SRT method as a suitable tool for measuring nonnative speech intelligibility. As the examples in section 4 show, the method is effective in collecting quantitative data for nonnative talkers as well as listeners. The coherent performance across-languages makes the method suitable for various multi-lingual applications.

6. References

[1] Flege, J.E. "The intelligibility of English vowels spoken by British and Dutch talkers," In *Intelligibility in Speech Disorders,* edited by R.D. Kent, John Benjamins publishing company, Amsterdam, 1992.

[2] Strange, W. "Cross-Language Studies of Speech Perception; A historical review," In *Speech Perception and Linguistic Experience,* edited by W. Strange, York Press, Baltimore, 1995.

[3] Flege, J.E. "Second-language speech learning: theory, findings, and problems," In *Speech Perception and Linguistic Experience,* edited by W. Strange, York Press, Baltimore, 1995.

[4] Flege, J.E., Bohn, O-S., and Jang, S. "Effects of experience on nonnative speakers' production and perception of English vowels," *J. Phonetics*, Vol 25: 437-470, 1997.

[5] Mayo, L.H., Florentine, M. and Buus, S. "Age of second-language acquisition and perception of Speech in noise," *J. Speech Lang. Hear.Res.* Vol 40, 686-693, 1997.

[6] Wijngaarden, S.J. van, Steeneken, H.J.M. and Houtgast, T. (2001). The effect of a non-native accent in Dutch on speech intelligibility. *J. Acoust. Soc. Am.* Vol. 109 (5) Pt 2: page 2473, 2001.

[7] Lane, H. "Foreign accent and speech distortion," *J. Acoust. Soc. Am*. Vol. 35: 451-453, 1963.

[8] Gat, I.N. and Keith, R.W. "An Effect of Linguistic Experience; Auditory Word Discrimination by Native and Nonnative Speakers of English," *Audiology* Vol. 17: 339-345,1978.

[9] Nábělek, A.K. and Donahue, A.M. "Perception of consonants in reverberation by native and nonnative listeners," *J. Acoust. Soc. Am*. Vol. 75:632-634, 1984

[10] Wijngaarden, S.J. van. "Speech intelligibility of native and non-native Dutch speech". *Speech Communication*, Vol. 35:103-113, 2001.

[11] Wijngaarden, S.J. van and Steeneken, H.J.M. "The Intelligibility of German and English Speech to Dutch Listeners". *Proceedings of the International Conference on Spoken Language Processing (ICSLP2000),* 2000.

[12] Bradlow, A.R. and Pisoni, D.B. "Recognition of spoken words by native and nonnative listeners: talker-, listener- and item-related factors," *J. Acoust. Soc. Am*. Vol. 106: 2074-2085, 1999.

[13] Plomp, R. and Mimpen, A.M. "Improving the Reliability of Testing the Speech Reception Threshold for Sentences", *Audiology*, Vol. 18: 43-52, 1979.

[14] Buus, S., Florentine, M., Scharf, B. and Canevet, G. "Native, French listeners' perception of American-English in noise." *Proceedings of Internoise 86*, 1986, pages. 895-898.

[15] Wijngaarden, S.J. van and Steeneken, H.J.M. "A Proposed Method for Measuring Language Dependency of Narrow Band Voice Coders". *Proceedings Eurospeech 2001, Aalborg, Denmark,* 2001.

[16] Wijngaarden, S.J. van. *Multilingual SRT 2.41. Database acquisition and speech reception threshold measuring program*. Soesterberg, TNO Human Factors, 2001.

EVALUATION OF SPEAKER'S DEGREE OF NATIVENESS USING TEXT-INDEPENDENT PROSODIC FEATURES

Carlos Teixeira[1,2], *Horacio Franco*[2], *Elizabeth Shriberg*[2], *Kristin Precoda*[2], *Kemal Sönmez*[2]

[1]IST/INESC-ID, Lisbon, Portugal
[2]SRI International, Menlo Park, CA 94025
carlos.teixeira@inesc-id.pt

Abstract

Giving feedback on the degree of nativeness of a student's speech is an important aspect of computer-aided language learning. This task has been addressed by many studies focusing on the segmental assessment of the speech signal. To better model human nativeness scores, other aspects of speech should also be considered, such as prosody. This study examines the use of prosodic information to evaluate the degree of nativeness of student pronunciation, independent of the text. Supervised strategies based on human grades are used in an attempt to select promising features for this task. Previous results obtained with non-native speakers showed improvements in the correlation between human and automatic scores. New strategies were evaluated with tests including native and non-native speakers. Specific features based on durations, namely for intra-sentence pauses, revealed potential use for further improvements.

1. Introduction

The aim of this work is to examine the use of prosodic information in evaluating the degree of nativeness of pronunciation for a text-independent task. This task has been addressed by many studies focusing on the segmental assessment of the speech signal [1, 2, 3, 4]. Recently, several studies have used suprasegmental speech information for computer-assisted foreign language learning (e.g. [5]). The present work's contribution is to attempt to select promising features, using a supervised selection strategy based on human scores of nativeness. While we expect prosody to carry information about the degree of nativeness of both sentences and individual words, in this study we concentrate on effects at the word level. Our methodology was based on three steps:

1. Feature extraction. Durational and melodic information was obtained from each sentence in the form of
   - Time alignments, obtained with SRI's DECIPHER^{TM} hidden Markov model (HMM) based speech recognition system [6]
   - Stylized pitch contours, from a model of dynamic prosodic information [7]

   Potentially useful and meaningful features were derived from this information and combined with lexical information.
2. Prosodic modeling. Decision trees were used to produce the automatic nativeness scores. These trees were generated using the same procedures and parameters as in previous studies [1]
3. Combination with other knowledge sources. The prosodic features used in this work were combined with previously computed scores of the degree of nativeness — based on spectral match and timing information [2] — in order to achieve higher correlations with scores given by human listeners.

Preliminary results with non-native speakers have shown improvements in the correlation between human and automatic scores [8]. These results are now augmented with test sets that include native speakers to provide a wider range of scores as well as a richer database for the calibration of nativeness scores.

2. Speech data and scoring

The corpus contained nearly a hundred adult native Japanese speakers. The set of speakers was fairly balanced on the basis of gender and English pronunciation abilities, which ranged from beginning to advanced. Each speaker read 145 sentences taken from a pool of 12,000 different English sentences. These included sentences from news broadcasts, literature, children's literature, and simple sentences written expressly for this use. In addition, a subset of the Wall Street Journal (WSJ) speech corpus was selected. This allowed our system to score the higher degree of nativeness for native speakers. The training part of this subset was also used to normalize some of the features from both native and non-native corpora.

2.1. Human scoring

Each utterance from the non-native speech corpus was graded by seven native American English speakers. The ratings were on a scale from 1 to 5, where a rating of 5 indicated very good pronunciation, and a rating of 1 indicated that the utterance had a strong foreign accent. The average correlation between the raters was computed to be 0.8 [1]. The median of the ratings from all raters was found for each utterance. A score of 6 was assigned to the utterances selected from the WSJ (native speakers). These values were used as the reference human scores and served as the inputs for the supervised classification approach used in this study.

2.2. Output of machine scores

Decision trees provide scores that can be evaluated by different measures of performance [3]. When the goal is to find a discrete score, as was asked of the human listeners, the highest posterior probability overall possible discrete scores (h_i) given

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

the machine score $\bar{m}$ can be used:

$$\tilde{h}_{opt} = \arg \max_{i \in [1,..,G]} [P(h_i|\bar{m})] \quad (1)$$

where G is the number of distinct grades.

A continuous score can also be derived. According to the minimum error criterion the optimal score is given by

$$E[h|\bar{m}] = \sum_{i=1}^{G} h_i \cdot P(h_i|\bar{m}). \quad (2)$$

2.3. Evaluation of machine scores

Two measures of performance were used on both discrete and continuous scores: the correlation and the error between the human and the automatic scores. This error is the average of the absolute value of the differences between the two scores. It is presented here as a percentage of the maximum error (difference between the highest and the lowest score of the scale used by the human listeners, i.e., 5).

3. Feature extraction

Many of the features are averages of measurements taken over the time. The remainder resulted from events that were uniquely defined in each utterance, such as the maximum or minimum of a feature. Gender was the only feature assumed to be known and the only one clearly based on specific speaker characteristics. Most of the features proposed are based on durations, normalized by the rate of speech (ROS) [4], which was itself used as a feature. The phone durations used were further normalized by the average phone durations estimated from a native English corpus (WSJ).

To define features related to prosody, we estimated a time instant for the primary stress in each word. These instants were then used as references for providing text-independent information. Three definitions of the time of primary stress were computed:

- The center of the longest vowel within each word, according to segmental forced alignments
- The center of the vowel carrying primary lexical stress
- The instant of time of maximum F0 excursion within each word, the nearest vowel to this instant was taken to be the primary stressed vowel

Using each of these definitions we computed three features that we refer to as the *word stress* features: duration of the assumed primary stressed vowel, duration between the center of this vowel and the center of the next vowel within the word, and duration between the center of the assumed primary stressed vowel and the center of the previous vowel within the same word.

3.1. Features derived from forced alignments

The following features are average durations, computed only with the information provided by the Viterbi forced alignments. We used averages of the duration of intra-sentence pauses, time between these pauses, and duration of words, vowels, and time between the centers of vowels. A subset of the WSJ corpus was used to compute the average native duration for each vowel in the phone inventory. The duration of each vowel in the utterance was normalized by the corresponding native average and used as a feature. Within each word the longest vowel was found and the word stress features were computed.

The lexical primary stressed vowel of each word was located in the forced alignments. Using this vowel, the word stress features and the duration to the next lexically stressed vowel (in a following word) were computed. This last feature represents an approach to estimating rhythm. The average time difference between the maximum F0 excursion and the longest vowel in the word completed this set of lexical features. These features were averaged over all words containing lexical primary stress in the utterance.

3.2. Features based on the pitch signal

The maximum F0 excursion within the utterance was taken as a feature [8]. The maximum and the minimum values for the pitch slope were found within each utterance and used as features. Based on pitch slope, each frame was also categorized as unvoiced, rising, or falling. Using these categories as a stream of symbols, a bigram was estimated for each utterance. The corresponding relative frequencies of transitions between categories were used as features. The number of rising frames before the maximum F0 excursion, and the number of falling frames after this instant, were both used as features. The number of changes in slope per frame was considered another feature attempting to capture the pitch variation.

We also computed the average duration of rising regions and the fraction of time these occupied within the utterance. The maximum duration of consecutive rises was computed as well as the increase in pitch inside this rising region. Similar features were computed for the falling frames. The ratio of the number of pitch rises to the number of pitch falls was also computed.

3.3. Features based on alignments and pitch information

Combining the information contained in the forced aligned transcriptions with the pitch information enables us to find the instant of maximum F0 excursion within each word and to measure time between this instant and other speech events found in the alignments. These measurements were then averaged for all the words in the utterance. This set of features included the value of the maximum F0 excursion, the time between the maximum F0 excursion and the center of the nearest vowel, the time between the maximum F0 excursion and the center of the longest vowel in the word, and the word stress features considering the maximum F0 excursion as the location of primary stress.

3.4. Features from unique events

Most of the features previously described are averages of events that can occur several times in the utterance. These kinds of features are more reliable for a text-independent approach; however, some unique events can convey important information about the degree of nativeness of an utterance. Three types of events were considered: two longest within-sentence pauses, two longest words, and two longest vowels within the utterance. The durations of each of these were taken to be features. For the two longest words we also measured the word stress features associated with the three different methods for defining the instant of primary stress.

4. Results and discussion

Previous experiments [8] performed with non-native speakers were repeated, including the subset of the native WSJ corpus. A few of the results from these experiments are represented in Table 1. These experiments aim to distinguish the performance of features based on segmental information (o2) from performance obtained just with the pitch signal (q2). We considered as segmental information (o2) the three base features (posterior, duration, and ROS scores, as proposed and evaluated in [1]) together with all the new features that do not use pitch or lexical stress information. In (p2) lexical-stress-based features were combined with segmental features (o2). The last experiment includes all the features described in this paper (r2).

	discrete scores		continuous scor.	
Features	corr.	error	corr.	error
(n2) 3 base features	0.732	14.3	0.763	14.6
(o2) segmental	0.743	13.5	0.767	14.3
(p2) + lexical	0.733	13.7	0.762	14.4
(q2) suprasegmental	0.272	23.6	0.321	22.3
(r2) all the above	0.728	14.0	0.763	14.5

Table 1: *Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

The use of the segmental features (o2) provided the best result. The improvements found in correlation, relative to the features used in previous studies (n2), are 1.5% for the discrete scores (3.4% with the non-native corpus) and only 0.5% for continuous scores (1.4% for only non-natives). As before, combining lexical primary stress information did not improve performance (p2). The use of all our suprasegmental features (q2) provides little information about the degree of nativeness. Finally, combining these features with segmental features (r2) did not lead to an improvement over using only the segmental features (o2). The results presented in Table 1 confirm previous conclusions [8]. In the following experiments we decided to follow a data-driven method for selecting a good set of features, instead of comparing results from categorical sets of features (e.g., segmental versus suprasegmental).

A first approach was based on the selection of the most successful single features in terms of continuous correlation. ROS (g2), duration (j2) and posterior (k2) scores, and average duration between intra-sentence pauses (t) have presented a continuous correlation higher than 0.4. These results are in Table 2. As in previous studies the posterior scores (k2) proved to be the more effective for the present task.

	discrete scores		continuous scor.	
Features	corr.	error	corr.	error
(g2) ROS	0.371	23.3	0.440	21.3
(j2) duration	0.445	21.0	0.511	20.2
(k2) posterior	0.700	15.8	0.730	15.6
(t) between pauses	0.407	20.7	0.427	22.2
(u) all the above	0.731	14.3	0.763	14.7

Table 2: *First feature selection approach. Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

The average duration between intra-sentence pauses (t) alone produces results comparable to previously derived features ROS and duration. The histograms for each of the given scores, of the values measured for this feature, are represented in Figure 1. It is clear from this figure that natives seldom speak continuously during a period as short as 50 to 100 ms, while non-natives do it more often as their degree of nativeness decreases. On the other hand, natives seems to be more confident about talking without any recognized pause during periods longer than 300 ms, while non-natives hardly do it for more than 250 ms.

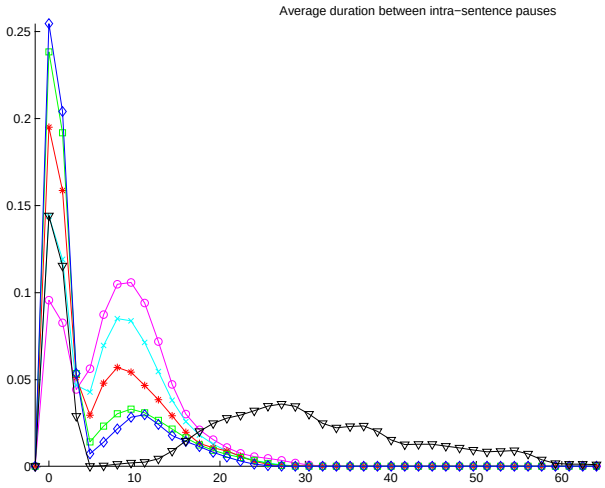


Figure 1: *Histograms of the average duration between intra-sentence pauses. Each histogram represents a single score value:* $\circ = 1, \times = 2, * = 3, \square = 4, \Diamond = 5, \nabla = 6$ *(native). The horizontal axis is in number of frames.*

Combining the single features, presenting continuous correlation higher than 0.4 (u), provides a result comparable to the use of all the features available (r2). However, this result does not show an improvement over previous studies (n2) and is not as good as the one obtained by using all the derived features based on segmentals (o2).

Table 3 represents some results obtained while following our second approach for achieving better scores while identifying additional relevant features. This approach makes use of the three base features in association with each one of the new features proposed in [8]. The results presented were selected from the experiments that have shown a continuous correlation of at least 0.765. The additional features used in these experiments were the average duration between lexically primary stressed vowels (v), average duration between the center of the longest vowel within the word and the center of the lexically primary stressed vowel (w), maximum pitch slope within the utterance (x), duration of the longest intra-sentence pause (y), duration of the second-longest intra-sentence pause (yy), longest word duration within the sentence (z), and relative frequency of the rising pitch frame followed by a falling pitch frame (zz).

The use of the longest intra-sentence pause (y) gives us an increase in the discrete and continuous correlation score of 2.5% and 0.7%, respectively, when compared with the use of the base features (n2). When compared with our previous best result (o2), these scores are only 0.9% and 0.1% better. However, the discrete correlation score of 0.75 is still the best ever found in this study. It is interesting to notice the small improvements found in experiments (x) and (zz), since the added features are based exclusively in the pitch information.

	discrete scores		continuous scor.	
Features	corr.	error	corr.	error
(n2) base features +	0.732	14.3	0.763	14.6
(v) btw_lex_stress&	0.740	13.7	0.766	14.2
(w) & max_vow	0.730	14.3	0.767	14.4
(x) max_F0_slope	0.730	14.2	0.765	14.5
(y) 1st_max_pause	0.750	13.6	0.768	14.3
(yy) 2nd_max_pause	0.739	14.1	0.766	14.5
(z) 1st_max_word	0.736	14.2	0.767	14.4
(zz) F0_rise+fall	0.735	14.2	0.766	14.5

Table 3: *Second feature selection approach. Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

Extending the principle of the first approach, allowing more than four features to be used together, we selected all features that, when used alone, provided a continuous correlation value higher than 0.10 (aa), 0.20 (ab), and 0.25 (ac) values. The more relevant results obtained for this third approach are in Table 4.

With this approach, the best continuous correlation was achieved in experiment (ab) where the selected features were: posterior and duration scores, ROS, average duration between intra-sentence pauses, duration of longest intra-sentence pause, duration of second-longest intra-sentence pause, second-longest word duration within the sentence, average duration between the center of the longest vowel within the word and the center of the lexically primary stressed vowel, maximum duration speech segment within which all frames had falling pitch, and relative frequency of a rising pitch frame followed by an unvoiced frame.

Features	discrete scores		continuous scor.	
cont. corr. $>$	corr.	error	corr.	error
(aa) 0.10	0.747	13.6	0.765	14.4
(ab) 0.20	0.740	13.7	0.769	14.3
(ac) 0.25	0.726	14.2	0.763	14.5

Table 4: *Third feature selection approach. Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

In the fourth approach, extending the principles of the second and third approaches, we selected all features that, as a result from the second approach, provided a continuous correlation value higher than 0.763 (ad) and 0.765 (ae) values. The higher correlation scores obtained are in Table 5. The second approach gave us good results, using only four features in each experiment. In the fourth approach we combined the features that provided the best results obtained with the second approach. However, this approach did not lead to a better performance than the second approach. On the contrary, the results are even slightly worse.

Features	discrete scores		continuous scor.	
cont. corr. $>$	corr.	error	corr.	error
(ad) 0.763	0.736	13.7	0.765	14.2
(ae) 0.765	0.735	13.6	0.764	14.3

Table 5: *Fourth feature selection approach. Correlation and error (%) between human and machine scores obtained with a corpus including both native and non-native speakers.*

In an earlier study [8], experiments made exclusively with non-native speakers did not lead to any improvement when pitch information was used in addition to the remaining proposed segmental features. This was also basically found in the experiments described in this paper, which also included a set of native speakers, apart from results (x) and (zz) in Table 3. On the other hand, improvements may be obtained from adding further specific features derived from the forced alignments. Some features based on durations — namely intra-sentence pauses — revealed potential use for improvements. We expect to continue this work in different directions. Future steps will include experiments investigating the performance of these features in discriminating between native and non-native speakers and further feature analysis and alternative supervised classification techniques.

5. Acknowledgments

We express our gratitude to Colleen Richey and Harry Bratt for their help. This work was supported by DARPA Agreement DASW01-96-3-0001 and NSF STIMULATE IRI-9619921. The views expressed here do not necessarily reflect those of the U.S. Government. The participation of the first author was partially supported by POSI E.U. Third Framework Programme.

6. References

[1] H. Franco, V. Abrash, K. Precoda, H. Bratt, R. Rao, J. Butzberger, R. Rossier, and F. Cesari, "The SRI $EduSpeak^{TM}$ System: Recognition and pronunciation scoring for language learning," (to appear) *Proc. of Integrating Speech Technology in Language Learning*, 2000.

[2] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," *Speech Communication*, vol. 30, pp. 121–130, 2000.

[3] H. Franco and L. Neumeyer, "Calibration of machine scores for pronunciation grading," *Proc. Int'l Conf. on Spoken Language Processing*, 1998.

[4] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic text-independent pronunciation scoring of foreign language student speech," *Proc. Int'l Conf. on Spoken Language Processing*, pp. 1457–1460, 1996.

[5] R. Delmonte, "SLIM prosodic automatic tools for self-learning instruction," *Speech Communication*, vol. 30, pp. 145–166, 2000.

[6] V. Digalakis and H. Murveit, "GENONES: Optimizing the degree of mixture tying in large vocabulary HMM based speech recognizer," *Proc. ICASSP 1994*, pp. 1537–1540, 1994.

[7] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," *Proc. Int'l Conf. on Spoken Language Processing*, 1998.

[8] C. Teixeira, H. Franco, E. Shriberg, K. Precoda, and K. Sönmez, "Prosodic features for automatic text-independent evaluation of degree of nativeness for language learners," *Proc. ICSLP*, 2000.

ADAPTATION METHODS FOR NON-NATIVE SPEECH

Laura Mayfield Tomokiyo and Alex Waibel

Language Technologies Institute
Carnegie Mellon University

ABSTRACT

LVCSR performance is consistently poor on low-proficiency non-native speech. While gains from speaker adaptation can often bring recognizer performance on high-proficiency non-native speakers close to that seen for native speakers [12], recognition for lower-proficiency speakers remains low even after individual speaker adaptation [2]. The challenge for accent adaptation is to maximize recognizer performance without collecting large amounts of acoustic data for each native-language/target-language pair. In this paper, we focus on adaptation for lower-proficiency speakers, exploring how acoustic data from up to 15 adaptation speakers can be put to its most effective use.

1. INTRODUCTION

As speakers learn a new language, they trace unique paths through acquisition of phonology, vocabulary, grammar, pragmatics, and even social aspects of spoken communication. The variability that this complexity engenders poses a serious problem for speech recognition. Once speakers reach a certain level of proficiency, their pronunciation may become fossilized, with the most noticeable features of their accent influenced by their native phonological system. In the early stages of learning, however, speakers experiment with new sounds, which results in phonetic realizations that are inconsistent and often distant from both the target phone and any native language phone that one might expect to influence it.

Study of the nature of non-native speech has suggested that perception of a phoneme is influenced by the phonetic contrasts that are meaningful in the speaker's native language (L1) [6] and that production is related to perception for allophonic contrasts [5]. However, it has also been observed that articulation of target language (L2) phones cannot be reliably traced to a related, interfering phone in the speaker's native language [1]. It has often been an assumption in efforts to adapt to non-native pronunciation in speech recognition that a speaker's realization of an L2 phone will fall "somewhere between" the average native realization and realization of an L1 phone that the speaker perceives as being similar to it. While adaptation based on this assumption has been successful for high-proficiency speech and simple tasks (e.g., [8, 14, 12],), both the speech degradation due to high cognitive load and the variability in articulation discussed above make recognition of lower-proficiency speech in LVCSR tasks a very hard problem [2].

In this paper, we concentrate on a specific group of lower-proficiency speakers, quantifying characteristics of their speech and comparing methods of adapting to it in LVCSR. Working with a controlled group of native speakers of Japanese, we investigate phonological properties of speech, fluency and disfluency, and reading errors in a read news task. We then discuss the effectiveness of training and of mixed-style and MLLR adaptation to the non-native condition, examining the contribution of L1 and L2 data to the adaptation process.

2. DATA

In this section, we describe the language background and proficiency evaluation of the speakers, the task and recording conditions, and the recognition system used for adaptation experiments.

2.1. Target speakers

The speakers in this study were all native speakers of Japanese. All had had 6-8 years of formal study of English and had lived in an English-speaking country for 6-12 months. All reported difficulty in making themselves understood, and rated their confidence in conversational speaking between 1.5 and 2.5 on an informal scale of 0 to 4. These speakers can be described as having a good grasp of the formal properties of English but limited productive ability.

In addition to informal evaluations, speaker proficiency in the test set was controlled with respect to scores on the formal SPEAK assessment [13]. All test speakers scored between 1.89 and 2.17 on the read speech portion of this test, which gives scores on a scale of 0 to 3 for identifiably non-native speech. Speakers assigned to the training set ranged from 1.44 to 2.83.

There were 10 test speakers, 15 training speakers, and 8 native speakers in this database.

2.2. Task

Two sets of speakers were recorded for this research. The primary group of interest, which included all test speakers, was recorded speaking English. A second group of speakers was recorded speaking their native language of Japanese.

2.2.1. Accented L2 data

Accented data, that is, recordings of native Japanese speakers speaking English, is referred to as *L2 data* because English is the speakers' L2.

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

Speakers completed a read news task in which they read aloud three articles from a children's news archive. This task was designed to mirror well-known tasks such as Wall Street Journal, which was determined during preliminary data collection to be too difficult for our speakers.

Of the three articles, one was common to all speakers and the other two were unique to each speaker. Article length averaged 50 sentences. The training/adaptation set represented approximately 3 hours of acoustic data.

Recording was done in a quiet room using a close-talking headset and a DAT recorder. Speakers were alone in the room while recording.

2.2.2. L1 data

Native-language data, that is, recordings of native Japanese speakers speaking Japanese, is referred to as *L1 data* because Japanese is the speakers' L1.

The L1 data that was used for model adaptation and training was taken from the Globalphone database [11] and consists of recordings of native Japanese speakers reading news articles from the Nikkei Shimbun in Japanese. Although the content of this newspaper is more difficult than that in children's news, the reduced cognitive load required for reading one's native language means that the difficulty of the L1 and L2 tasks was similar for the native Japanese speakers.

Speakers recorded an average of 15 minutes of speech. Recording was done in a quiet room using a close-talking headset and a DAT recorder. For consistency with the accented L2 data, 3 hours of this speech distributed across 15 speakers was used for training and adaptation.

2.3. Recognition system

All experiments described in this paper used the JRTk speech recognition toolkit [4] with fully continuous context-dependent acoustic models and a trigram language model. Context-dependent models were determined experimentally to perform better than context-independent models for this speaker set and task. Vocal tract length normalization and cepstral mean subtraction are applied at the speaker level. Linear discriminant analysis (LDA) is used to find the most discriminative of the MFCC, delta, and power features and reduce the dimensionality of the feature vector describing each frame. WER figures always represent accuracy after speaker-dependent MLLR adaptation on 50 utterances. Performance of this system on Broadcast News F0-condition speech is 9.4%. Because of differences in speaking style (informal vs. professional anchor) and language modeling (the broadcast news model was adapted to children's news, but is still not optimal for the task), performance on local native speakers on the children's news task is significantly higher, at 19.2%.

3. CHARACTERIZING LOW-PROFICIENCY ENGLISH

Learning to speak a new language is a journey that doesn't always follow a straight line from L1 to L2. For many speakers, reaching proficiency is a matter of years of trial and error. In this section, we discuss some of the features of non-native speech of the proficiency level we are targeting.

3.1. Reading errors

Reading errors, which are commonly assumed not to occur often enough to greatly affect system performance, were frequent in our data. Nearly 3% of the words that were read by the non-native speakers were not the words on the page, as compared to 0.4% for native speakers.

In addition, the types of reading errors that were made were distributed quite differently in native and non-native speech. Substitution of a morphological variant was by far the most common reading error in non-native speech. Singular-plural substitution represented over 60% of these morphological errors. Non-native reading errors were more likely to affect the syntactic integrity of the sentence; for example, the sentence "Doctors are studying the pill's *effect* on patients" is meaningful whether the word *effect* is singular or plural, whereas the sentence "American *student* perform poorly on standardized tests" is made syntactically incorrect by the speaker's substitution of *student* for *students*. A more detailed breakdown of reading errors in this data can be found in [9].

3.2. Phonological properties

A segment of the non-native data collected in this project was phonetically transcribed by experienced transcribers. Although a number of expected transformations (e.g., /ɪ/→ [i]) were verified during this process, the principal observation was that the number of realizations that could not be transcribed using the union of the standard American English and Japanese phone sets was great. Transcribers required an extensive set of supplemental diacritics, representing r-coloring, centering, and palatization, among other things, to begin to capture the data. There was also a great deal of intra- and inter-speaker inconsistency. One speaker, for example, consistently pronounced [ʌ] as [ɚ] – but only in the second half of one article. For some reason, he made the decision to try this pronunciation out, and then abandoned it when he began the next reading.

Diversions from standard American English phonology were also found in recognizer-driven analysis. Phoneme-level recognition of the data revealed both common insertions, deletions, and substitutions and high overall levels of phoneme confusion, consistent with observations from manual analysis. In an experiment designed to uncover lexical variants, it was found that when phone-level insertions, deletions, and substitutions are considered, 57% of the polyphones (5-phone sequences) in the test data were not seen in the training data, compared to 92% for native speech.

3.3. Fluency

The low-proficiency speakers targeted in this paper read far more slowly and haltingly than native speakers do. Frequent inter-word pauses, stumbling over words, and multiple repetitions of sequences of words have implications for both acoustic and language modeling. In particular, it has been our experience that no complex cross-word modeling is necessary for the lower-proficiency

speakers because words are usually articulated one at a time, with pauses in between them.

feature	mean		std. dev.	
	N	NN	N	NN
pause duration	9.56s	17.14s	3.16	7.33
phone duration	0.08s	0.12s	0.01	5.36
pause:word ratio	1:10	1:3	0.05	0.08
words/second	3.80	2.15	0.26	0.29
repair rate	0.57	2.25	0.33	1.42
repeat rate	0.07	0.34	0.07	0.23
retrace rate	0.58	2.35	0.35	1.26
retrace length	2.55	2.57	1.04	2.29
filler word rate	0.01	0.16	0.02	0.32
partial word rate	0.45	1.52	0.20	1.05

Table 1. Comparing fluency-related statistics for native (N) and non-native (NN) speakers in the reading task

Figure 1 gives statistics for fluency (and disfluencies) for the low-proficiency non-native speakers targeted in this paper. The non-native speech is clearly more disfluent than the native speech, as measured by such diverse features as speaking rate, ratio of silence to words, and number of repaired and abandoned words. The only feature that appears to be similar for native and non-native speakers is retrace length, or the number of words a speaker "rewinds" when correcting himself. It could be that this span is influenced by the syntax of the text, which is the same for both native and non-native speakers; it has also been suggested that retrace length is constant across languages [3].

4. SPEAKER PROFICIENCY AND RECOGNIZER PERFORMANCE

In this paper, we specifically target lower-proficiency speakers. Our premise is that these speakers may need processing different from that applied to higher-proficiency speech in order to raise recognition accuracy to an acceptable level. This assumption is based on the intuition that lower-proficiency speakers are somehow harder to understand, as well as the observation that these speakers are diverse and inconsistent in their articulation. To support our assumption, let us quantitatively examine the correspondence between proficiency and recognizer performance.

Figure 1 shows how word error rate (WER) varies with speaker proficiency. We see three distinct clusters. The cluster on the far right represents native speech; native speakers automatically receive a SPEAK score of 4. The center cluster represents speakers who scored between the test set cutoff of 2.17 (the lowest actual score in this group was 2.44) and the maximum non-native score of 3. The test speakers targeted in this paper fall into the leftmost cluster. Although there is some variation in recognizer performance within the clusters, speakers in the lower-proficiency group clearly are recognized with less accuracy than those in the other two.

5. ACOUSTIC MODEL ADAPTATION

In this section, we discuss offline adaptation to the non-native condition prior to individual run-time speaker

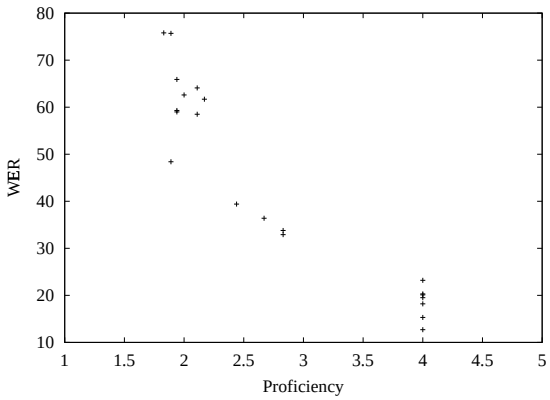


Fig. 1. Correspondence between speaker proficiency and recognizer performance in terms of word error rate (WER)

adaptation. We strive to answer two questions:

- Does L1 material provide better adaptation data than accented L2 data?
- Does mixed-style adaptation perform better than MLLR adaptation for non-native speech?

It was observed in [8] that MLLR adaptation with L1 LVCSR data gave similar improvements in accuracy to adaptation with accented L2 data when only isolated L2 phone data was available. In this paper, we explore a matched condition: same-domain LVCSR data is used for both L1 and L2 adaptation material.

In order to use the L1 data described in Section 2.2.2 for adaptation of English acoustic models, the Japanese lexicon had to be converted to the English phone set. Data-driven and IPA-based approaches to this problem have been studied (e.g. [11, 14, 8]); we used a combination. Pronunciation networks for each Japanese word were created with each Japanese phone replaced by a set of parallel transitions representing substitutions of related English phones and phone sequences. "Related" was defined to mean sharing all but one phonological feature. Therefore, any phone that differed only in place of articulation, or manner, or voicing, or vowel height, was added to the network. A forced alignment pass was then run on this network to find the path with the most likely match. Context-sensitive (considering preceding and following phone) global mappings were assigned based on the substitutions selected most often during alignment.

5.1. Mixed-style adaptation

In mixed-style training, adapted model parameters are estimated separately for each of the "styles" (in this case, L1 and L2), and then interpolated using a global interpolation weight. This is, in effect, a simple form of MAP adaptation, where an optimal weighting factor is determined experimentally rather than separately for each Gaussian based on the *a priori* distribution of the Gaussian parameters. It has been our experience that this method produces results that are similar to or slightly better than conventional MAP. If it is likely

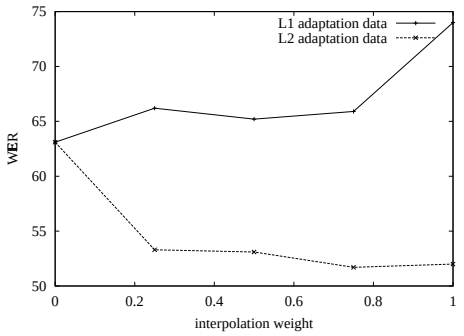


Fig. 2. Mixed-style adaptation using L1 and L2 adaptation data

	L1 data	L2 data
0 speakers (baseline)	63%	
3 speakers	68.1	58.1
15 speakers	73.4	52.5

Table 2. MLLR adaptation using varying amounts of L1 and L2 adaptation data (figures represent WER)

that the adaptation data represents the test data well, it can be heavily weighted for interpolation. As with MAP adaptation, this method performs better as the amount of adaptation data increases, as if individual parameters cannot be reliably estimated from sparse sample data no adaptation is performed. In this experiment, 15 adaptation speakers were used.

Figure 2 shows system performance after mixed-style adaptation with both L1 and L2 data. On the horizontal axis is the interpolation weight. When the interpolation weight is 1, the adapted mean is identical to the sample mean. When the interpolation weight is 0, the adapted mean is identical to the prior mean (i.e., there is no adaptation).

A clear degradation can be seen from adapting with L1 data, while the positive contribution of the accented L2 data can be seen rising steadily as the interpolation weight increases.

5.2. MLLR

In MLLR adaptation, transformation classes are defined, and model parameters of the entire class are shifted in the same direction. While this clustering allows MLLR adaptation to provide a general transformation with a small amount of adaptation data, there is a risk of shifting an individual parameter *away* from observed sample value, which is avoided in mixed-style adaptation.

Results of MLLR adaptation with L1 and L2 data are shown in Table 2. As with mixed-style adaptation, we see a degradation with the introduction of L1 acoustic material. The effect is more extreme with more adaptation speakers, indicating that sample means from the L1 data are not representative of the means in the actual accented test speech. Adaptation with accented L2 data, on the other hand, significantly improves performance over the baseline.

Results are given for 3 and 15 adaptation speakers. It is clear that the effectiveness of adaptation increases

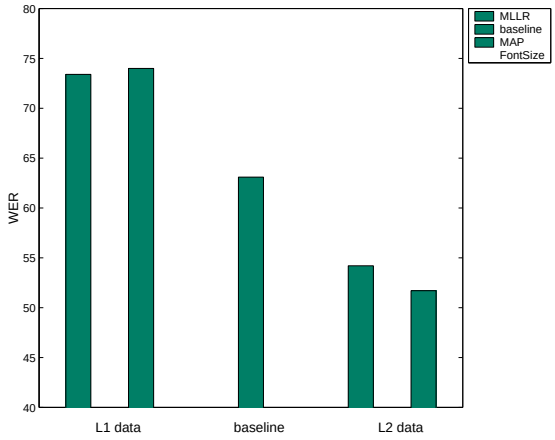


Fig. 3. Comparison of MLLR and MAP adaptation for 15 adaptation speakers

with the amount of adaptation speech. There are two reasons for this: more examples of sample values allow a more reliable estimate of the sample mean, and the more diverse set of samples contributes to a more general model.

5.3. Comparison of adaptation methods

Figure 3 contrasts MLLR and mixed-style adaptation performance for L1 and L2 adaptation material and 15 adaptation speakers. Both show similar trends, with mixed-style slightly outperforming MLLR.

We see clearly from all the experiments shown here that using L1 acoustic material for adaptation to low-proficiency non-native speech without re-evaluation of the polyphone set results in a degradation of recognizer performance, while adaptation with accented L2 data boosts performance.

6. RETRAINING WITH ACCENTED DATA

It was shown in Section 5 that while using accented data for adaptation improves recognition performance, adapting with L1 data results in a performance degradation. In speaker adaptation, the model inventory is kept the same, but the expectation of what a model sounds like is shifted towards what has been seen in the limited set of adaptation speech. The L1 data does not have the chance to make its maximal contribution, as the model inventory is based on the polyphones found in native speech; two allophones that are quite different in L1 may be used to update the same model if the two contexts do not trigger variation in English. By rebuilding the system based on the contexts that are meaningful in L1, we may be able to use the L1 data to its full advantage.

In this section, we compare systems trained with L1 data with systems trained with accented L2 data. Both full rebuilding of the system (rebuilding from scratch) and repetition of the final step of training (additional forward-backward iterations) with the new data are examined.

6.1. Rebuilding from scratch

In this experiment, two new systems were built, using L1 and accented L2 data. In both cases, initial labels were written using the baseline acoustic models, and a context-dependent system was trained along the specifications given in Section 2.3. Because the adaptation data available was sparse for fully training a recognizer, it was pooled with native English data in these experiments. The large amount of native data contributes to the robustness of the model, while the smaller amount of L1 or accented L2 data ensures that L1-specific phone sequences and phone realizations are seen during clustering and training. Training data consisted of 3 hours of L1 or accented L2 acoustic data pooled with the original native training data.

6.2. Additional forward-backward iterations

In this experiment, the new system was not retrained from scratch; rather, two additional forward-backward iterations are run on the fully trained baseline models using the accented L2 acoustic data.[1] In Section 5, we saw how recognition improves with adaptation to the non-native condition when accented data is used. By training with the accented data, we are essentially extending this approach, updating not only the mixture means but also the mixture weights and covariances. We also benefit from the second re-estimation. The effect of additional forward-backward iterations with the L1 data was not examined in this experiment.

6.3. Comparison of training methods

Figure 4 contrasts performance of fully-rebuilt and partially retrained systems. With the rebuilt systems, we see a small improvement when training with L1 data and a much larger improvement when training with accented L2 data.

The improvement from the additional training iterations is even larger. This may be because in retraining (described in Section 6.2), we are capitalizing on consistency in the data in the two phases of system building with native speech and retraining with non-native speech. When the two data sets are combined from the outset (as described in Section 6.1), we may incorporate a broader range of polyphones but be harmed by the mismatch between native and non-native speech. By simply retraining, we fix the identities of the acoustic models with native data, and then use the non-native data to adjust the expectation of how those models correspond to phonetic realization in non-native speech.

6.4. Model interpolation

Simply running additional forward-backward iterations with the three hours of accented data resulted in a 24% relative improvement over the baseline error rate. In this new model, however, the parameters were trained on a small amount of data. This introduces a danger of overfitting, a problem which can addressed by smoothing the models via interpolation with a more robust model [7]. A direct parameter interpolation technique has also been shown to be successful in creating context-independent non-native models from source and target language model sets [14]. In model interpolation experiments, it was our goal to move the retrained distribution back towards the native distribution to the point of maximum robustness.

In the interpolation method that we used, corresponding codebook weight, mean, and covariance matrix elements are linearly interpolated for each baseline system / retrained system acoustic model pair. This results in a covariance space that covers an area between the two original covariances, rather than the union of the two. We are able to interpolate the individual models in this way because there is a clear one-to-one mapping between models; the decision of which models to interpolate would be much more difficult if we were working with the re*built* system of Section 6.1 instead of the re*trained* system of Section 6.2. Our method is described in detail in [10]. Performance of the interpolated system is 29% above that of the baseline system, a significant improvement over the retraining alone. The effect of the interpolation weight on recognition accuracy is shown in Figure 5; optimal performance is found when the retrained models are weighted at .72.

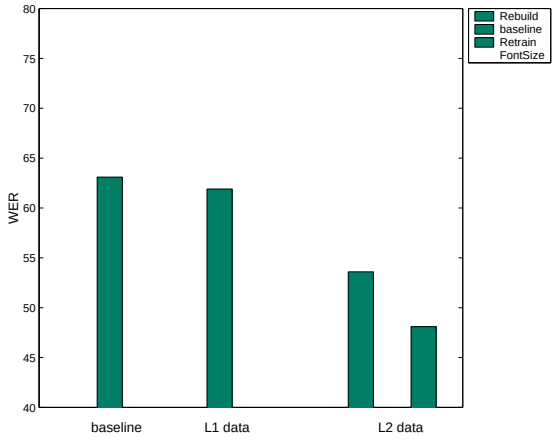


Fig. 4. Comparing rebuilding from scratch with L1 and L2 data and partial retraining with L2 data

7. SUMMARY

In this paper, we have examined how application of acoustic model training and adaptation techniques affects recognition accuracy on non-native speech. A summary of the individual contributions of each method is shown in Figure 6.

Generally speaking, adaptation to the non-native condition (and by adaptation we refer to both the speaker adaptation techniques of MAP and MLLR and retraining techniques) using L1 data does not improve performance, and in some cases causes a large degradation. Accented L2 data, on the other hand, contributes positively to the acoustic model. The largest gains are seen when using the full 3 hours of accented data to run ad-

[1] For this experiment, the baseline models were also trained an additional two iterations to ensure that the comparison was fair. We did not observe any significant change between the original 7-iteration training and the 9-iteration training with the native data, however.

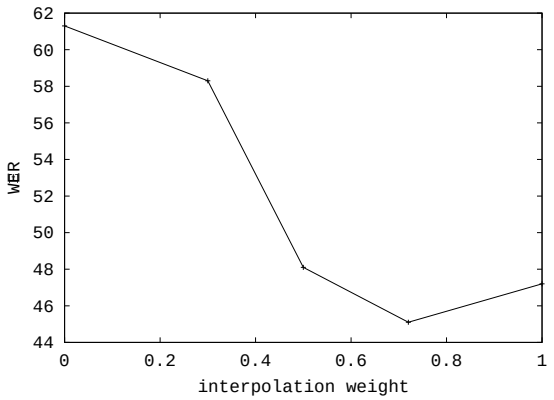


Fig. 5. Results for interpolation with different interpolation weights. A weight of 0 represents performance with the original acoustic models. A weight of 1 represents performance with the new models.

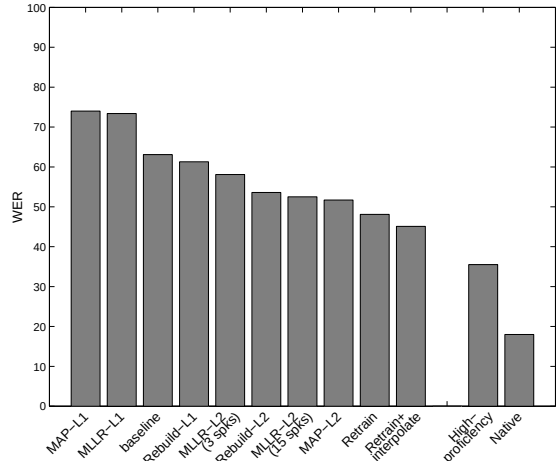


Fig. 6. Summary of adaptation results. Bars labeled "MAP" refer to mixed-style adaptation, which is a simplified form of MAP adaptation.

ditional forward-backward training iterations and then interpolating the retrained model back with the more robust baseline models. Significant gains are also seen with MAP and MLLR adaptation, where performance of the system improves proportionally to the amount of accented adaptation speech.

In the best case, word error rate for the lower-proficiency speakers is lowered from 63.1% to 45.1%, which represents a 29% relative reduction in error. This approaches, but does not match, performance on the higher-proficiency speakers.[2] With an absolute reduction in error of 18%, we have closed half of the gap in recognizer performance on native and low-proficiency non-native speech; how close this brings us to the upper limit, however, remains to be seen.

[2]We see the same trends when applying adaptation techniques to proficient non-native speech, although the effect is far less dramatic [10].

8. REFERENCES

[1] Eugene Brière. An investigation of phonological interference. *Language*, 42(4):768–796, 1966.

[2] William Byrne, Eva Knodt, Sanjeev Khudanpur, and Jared Bernstein. Is Automatic Speech Recognition Ready for Non-Native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English. In *Proc. Speech Technology in Language Learning (STiLL)*, 1998.

[3] Robert Eklund and Elizabeth Shriberg. Crosslinguistic Disfluency Modeling: A Comparitive Analysis of Swedish and American English Human-Human and Human-Machine Dialogs,. In *Proc. ICSLP*, 1998.

[4] Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries, and Torsten Zeppenfeld. The Janus-RTk Switchboard/Callhome 1997 Evaluation System. In *Proc. the LVCSR Hub5-e Workshop*, 1997.

[5] James Emil Flege. Production and perception of a novel, second-language phonetic contrast. *J. Acoust. Soc. Am.*, 93(3):1589–1608, March 1993.

[6] Robert Allen Fox and James Emil Flege. The perception of English and Spanish vowels by native English and Spanish listeners. *J. Acoust. Soc. Am.*, 97(4):2540–2551, April 1995.

[7] X.D. Huang, Mei-Yuh Hwang, Li Jiang, and Milind Mahajan. Deleted interpolation and density sharing for continuous hidden markov models. In *Proc. ICASSP*, Atlanta 1996.

[8] Wai Kat Liu and Pascale Fung. MLLR-based accent model adaptation without accented data. In *Proc. ICSLP*, 2000.

[9] Laura Mayfield Tomokiyo. Handling Non-native Speech in LVCSR: A Preliminary Study. In *Proc. Incorporating Speech Technology in Language Learning (InSTIL)*, 2000.

[10] Laura Mayfield Tomokiyo. *Recognizing Non-native Speech: Characterizing and Adapting to Non-native Usage in Speech Recognition*. PhD thesis, Carnegie Mellon University, 2001.

[11] Tanja Schultz and Alex Waibel. Language independent and language adaptive large vocabulary speech recognition. In *Proc. ICSLP*, Sydney, 1998.

[12] Richard Schwartz, Hubert Jin, Francis Kubala, and Spyros Matsoukas. Modeling Those F-Conditions - Or Not. In *Proc. the 1997 DARPA Speech Recognition Workshop*, 1997.

[13] Guide to SPEAK. Produced by the Test of English as a Foreign Language Program, Princeton, NJ, 1987.

[14] Silke Witt and Steve Young. Offline Acoustic Modeling of Non-native Accents. In *Proc. Eurospeech*, 1999.

HMM-Based English Speech Recognizer for American, Australian, and British Accents

Rathi Chengalvarayan

Lucent Speech Solutions
Lucent Technologies Inc.
2000 Lucent Lane, Naperville
Illinois 60566, USA
rathi@lucent.com

Abstract

In this study, we focus on the problem of making a transition from the accent-dependent to accent-independent speech recognition technology in telephone communication devices. Previous studies showed that the multi-transitional model architecture is one of the solutions for robust speech recognition. In this paper, we investigate many universal hybrid systems that are trained with data recorded through Australian, American, and British accented speech for English language. This new universal system uses less than double the number of parameters as in individual system (American or Australian or British acoustic models) and significantly reduces the model parameters without affecting the performance when compared with multiple classifiers or multi-transitional models. We compare the performance of universal hybrid system on several independent connected-digit telephone test databases and demonstrate the effectiveness of hybrid architectures with data taken from all three regional accented speech.

1. Introduction

In recent years there has been an increased interest in approaches to speech and pattern processing that go beyond the conventional hidden Markov model (HMM) framework [4]. This has been motivated by limitations of current systems both in their error performance and by the fact that current statistical modling assumptions are inconsistent with that of natural speech signals [7]. It has been observed that current models are fragile in noise and are limited in their ability to handle pronunciation variations [14]. Speech recognition under accent variations is a challenging problem for which there are no completely satisfactory solutions [1]. This problem is crucial for the development of successful real-time multilingual applications in promising domains such as accent-independent speech recognition [17]. The speech for a particular language is rapidly changing depending on the regional accents [12]. Speech recognition suffers from significant performance deterioration when they are operated in mismatched accent conditions [2]. Collecting data in an accent-dependent environment is a key factor to understanding and solving accent problems [6].

Dialect also plays an important part in the overall degradation, resulting in different pronunciations for the same word [9, 20]. There are many ways to reduce the accent and dialect variations within a given language [21]. A typical approach is to integrate an accent classifier followed by a corresponding accent-specific recognizer [13]. Many systems can get 100% correct accent classification when tested on training data, but can get an average of 81% on the 10 sec test utterance [1, 10, 21]. It has been demonstrated in previous research efforts, that the multi-HMMs and multi-transitional architectures are many of the proposed solutions for robust recognition [16]. The idea is to provide more variability to the system to be trained, and to support this variability with the greatest number of parameters. The main drawback of utilizing the currently available systems is that the model size, and the computational complexity increases linearly related to different accents [5].

In this work, we propose several hybrid architectures that are trained with pooled data from American, Australian, and British accented speech. Experimental results on connected-digit recognition task show an average string error rate reduction of about 62% and 8% when compared to our best monolingual and multi-transitional systems respectively. The result indicates that the universal model is about three times faster and half time smaller than the multi-transitional or multilingual models and this makes it an ideal choice for practical accent-independent speech recognition applications.

2. Speech Database

This section describes the database, LL_US, used in this study [19]. This database is a good challenge for speech recognizers because of its diversity. It is a compilation of databases collected during several independent data collection efforts, field trials, and live service deployments. The LL_US database contains the English digits *one* through *nine*, *zero* and *oh*. It ranges in scope from one where talkers read prepared lists of digit strings to one where the customers actually use an recognition system to access information about their credit card accounts. The data were collected over network channels using a variety of telephone handsets. Digit string lengths

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

Data	Training		Testing	
	Strings	Speakers	Strings	Speakers
LL_AU	5298	800	848	200
LL_UK	2561	700	505	300
LL_US	7461	5234	2023	2512
Total	15320	6734	3376	3012

Table 1: Regional distributions of spoken digit strings and the speaker population among the training and testing sets of LL_AU, LL_UK, and LL_US databases.

range from 1 to 16 digits. The LL_US database is divided into two sets: training and testing. The training set includes both *read* and *spontaneous* digit input from a variety of network channels, microphones and dialect regions. The testing set is designed to have data strings from both matched and mismatched environmental conditions. All recordings in the training and testing set are valid digit strings, totaling 7461 and 2023 strings for training and testing, respectively. Only the 10, 14 and 16 digit strings were selected for testing.

The LL_UK consists of SpeechDat (M) database available through the European Language Resource Association [22]. This database was collected over the U.K. landline telephone network. Recordings were done using an ISDN telephone interface, yielding 8 KHz, 8-bit samples A-law coded signals. Each corpus contains the speech of 1000 speakers (about 500 male and 500 female). Most items are read, some are spontaneously spoken. Speech material is conveniently split into two disjoint sets, a training one and a testing one. The LL_UK database contains the third pronounciation for zero as *nought*, and multiple related word such as *double*. The training database consists of digit string lengths range from 1 to 16 digits that were spoken by 700 speakers (350 male and 350 female) for a total of 2561 valid strings. The testing database has 300 speakers (107 male and 193 female) and only the valid digit strings were selected for a total of 505 strings. Only the digit strings with length of 4, 10, 11 and 16 were chosen for testing.

The LL_AU consists of SpeechDat (II) database that was collected over the Australian landline telephone network [23]. This corpus contains the speech of 1000 speakers from all over the world. The training database consists of digit string lengths range from 1 to 16 digits that were spoken by 800 speakers for a total of 5298 valid strings. The testing database has 200 speakers and only the digit strings of length 5, 6, 10 and 16 were selected for a total of 848 strings. The LL_AU database contains the compact word *triple* in addition to LL_UK vocabulary. None of the speakers in the testing database appeared in the training databases. The data distribution of the training and testing set is shown in Table 1 for all three databases.

3. Robust HMM Architectures

When the accent of a particular language is unknown, the important mismatch between training data and signal encountered in recognition phase decreases drastically the performances of the recognition systems [17]. In this section, the HMM architecture of many different system configuration is discussed with great detail to reduce the accent and dialect variations within English language.

The recognizer feature set consists of 39 features that includes the 12 liftered linear predictive cepstral coefficients, log-energies, their first and second order derivatives. The energy feature is batch normalized during training and testing. Each feature vector is passed to the recognizer which models each word in the vocabulary by a set of left-to-right continuous mixture density HMM using context-dependent head-body-tail models. In this study, we model all possible inter-word coarticulation and each model is represented with 3 or 4 states, each having multiples of 4 mixture components. Silence is modeled with a single state model having 32 mixture components [19].

Training included updating all the parameters of the model, namely, means, variances and mixture gains using one iteration of K-means based maximum likelihood training procedure followed by six epochs of MSE training to further refine the estimate of the parameters [15]. Each training utterance is signal conditioned by applying batch-mode cepstral mean subtraction prior to being used in MSE training [5]. The number of competing string models was set to four, the step length was set to one and the length of the input digit strings is assumed to be unknown during the model training and a known-length grammar is used during testing. Penalties based on duration distributions are also applied to the likelihood score.

3.1. Baseline Acoustic Models

Three monolingual context-dependent head-body-tail digit models for American (US), Australian (AU) and British (UK) English accents were trained using data from the corresponding accent. Notice that the US model has 276 HMMs, the UK model has 304 HMMs and the AU model has 307 HMMs. The UK model has additional 28 context-dependent HMMs for the words *double* and *nought*. The AU model has three more HMMs for the word *triple* with silence contexts.

3.2. Multiple Acoustic Models

One of the multiple classifiers approach is employing three accent-dependent speech recognizers to decode the given input speech as shown in Fig. 1. The best candidate with a top likelihood score is chosen from the output of three parallel recognizers. We indicate this model as MULTI_LGL and is effective but rather expensive because the computation requirement is tripled [5]. Instead of picking the candidate with the best score, one can also pick the best hypothesis based on lower average-arc-count. We represent this model as MULTI_ARC. Notice that the MULTI_LGL and MULTI_ARC use the baseline

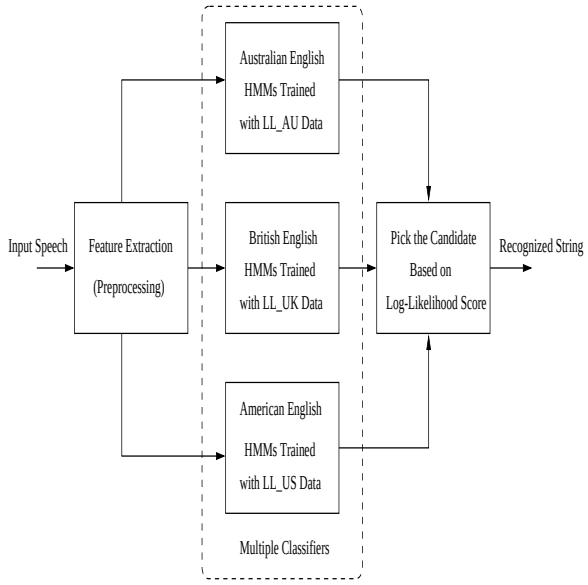


Figure 1: A block diagram of a MULTI_LGL parallel speech recognition system.

models for their decoding process.

For the purpose of comparison, multi-transitional (MULTI_PRN) and multi-grammar (MULTI_GRM) models were also created separately. Note that the multi-transition models were constructed by combining all the three monolingual models such that the decoder picks up the best model for a given utterence from an unknown accent source. The MULTI_PRN can use mix-and-match of digit models during recognition of a given digit string. For example, in a three-digit long string, the first digit can be derived from AU model, the second digit can be evolved from UK model, and the third digit can be generated from US model as shown in Fig. 2. We call this model as *multiple pronunciation*, since each digit has three different pronunciation or accent variability [3]. The multi-grammar model can follow only one optimized path out of the three possible combinations during recognition as illustrated in Fig. 3. The main difference between MULTI_GRM and MULTI_LGL is that the MULTI_LGL takes the matched silence model during the complete decoding path where as the MULTI_GRM sprinkles three different silence models whenever necessary in order to get an optimal decoding path. The MULTI_GRM model configuration seems to be more attractive than the other types of multiple classifiers due to its inherent flexibility.

3.3. Universal Acoustic Models

The multiple classifier approach can be cumbersome and will be difficult to handle more accent-specific utterances due to increased model complexity. This motivates the need for simple accent-independent universal hybrid system. This system uses a single decoder for British, Australian and American English digits, and is capable of recognizing digits with words from all accents as exem-

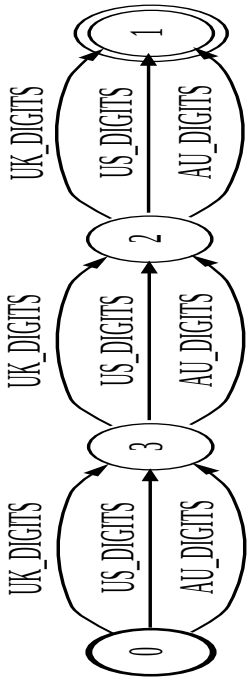


Figure 2: State diagram for the three-digit known-length grammar for MULTI_PRN.

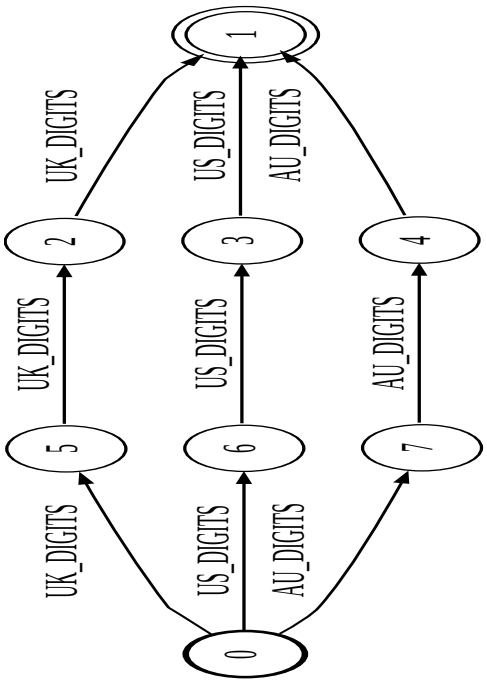


Figure 3: State diagram for the three-digit known-length grammar for MULTI_GRM.

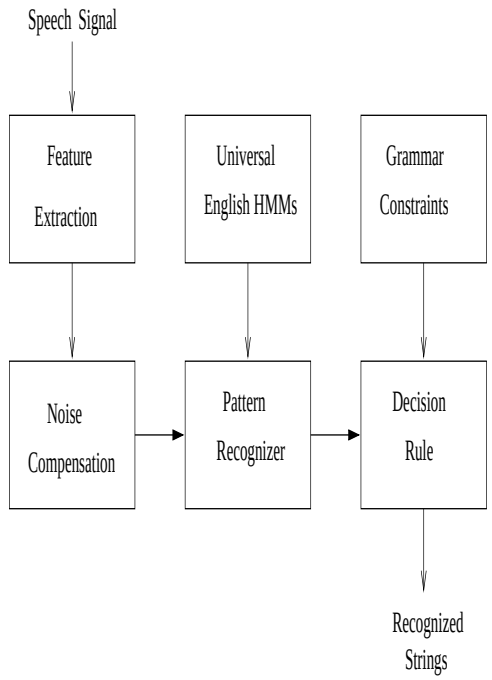


Figure 4: Block diagram of the HMM-based speech recognizer using universal modeling approach.

plified in Fig. 4. The acoustic models are trained using a pooled data recorded through Australian, American, British accented speech for English language and we name this model as UNIV_1. This training method is very similar to multi-conditional training, the whole system is trained using all available data, as reported in [11]. A negative side effect of this shared data is the increased possibility of confusion among words from three accents. This is overcome by doubling the Gaussian densities per state for the head and tails of a context-dependent head-body-tail topology. This newly expanded UNIV_2 topology is able to more adequately model the accent variations with increased set of HMM parameters [18].

4. Experimental Results

We have conducted experiments to compare the performance of universal hybrid system on several independent test databases and to demonstrate the effectiveness of a hybrid built with data taken from all three regional accented speech. The Table 2 through 4 presents the string accuracy and average-arc counts for three different monolingual models using all the three test datasets. When the LL_AU data is tested on a US system, the arc-count increases tremendously to a point where the recognizer is overloaded with unnecessary log-likelihood computations and arc-expansions, which result in a longer delay in reporting the recognized string. This is true to some extent that the US models look more fuzzier when tested on mismatched LL_AU data and hence the average-arc-count gets incremented and slow down the decoding speed. Similarly when the LL_US data is tested using AU

Data Model	LL_AU	
	String Accuracy	Arc Count
AU	90.2%	16791
UK	75.4%	20453
US	46.2%	21701
MULTI_PRN	90.3%	81750
MULTI_LGL	89.9%	58945
MULTI_ARC	87.2%	58945
MULTI_GRM	89.6%	34957
UNIV_1	88.3%	19127
UNIV_2	90.1%	20717

Table 2: String accuracy and arc-count for a known-length connected-digit recognition task using landline Australian English (LL_AU) data as a function of various model type.

Data Model	LL_UK	
	String Accuracy	Arc Count
AU	81.2%	17160
UK	90.5%	16107
US	62.4%	18230
MULTI_PRN	89.5%	67885
MULTI_LGL	90.3%	51497
MULTI_ARC	86.2%	51497
MULTI_GRM	90.1%	30226
UNIV_1	89.3%	17634
UNIV_2	90.3%	18257

Table 3: String accuracy and arc-count for a known-length connected-digit recognition task using landline British English (LL_UK) data as a function of various model type.

Data Model	LL_US	
	String Accuracy	Arc Count
AU	53.6%	30079
UK	64.5%	31094
US	93.6%	19675
MULTI_PRN	91.5%	88257
MULTI_LGL	90.3%	80848
MULTI_ARC	90.7%	80848
MULTI_GRM	93.1%	35683
UNIV_1	92.5%	23458
UNIV_2	93.1%	23818

Table 4: String accuracy and arc-count for a known-length connected-digit recognition task using landline American English (LL_US) data as a function of various model type.

Model Type	Model Size	String Accuracy	Average Arc Count
AU	1.54MB	75.0%	21343
UK	1.50MB	76.8%	22551
US	1.36MB	67.4%	19869
MULTI_PRN	4.40MB	90.4%	79297
MULTI_LGL	4.40MB	90.2%	63763
MULTI_ARC	4.40MB	88.3%	63763
MULTI_GRM	4.40MB	90.9%	33622
UNIV_1	2.60MB	90.1%	20073
UNIV_2	2.60MB	91.2%	20931

Table 5: Model size, average string accuracy and arc-counts across LL_AU, LL_UK and LL_US databases as a function of model type.

model the arc count increases due to mismatched testing data. The same observation can also be made for British accented data and the correponding UK model. The monolingual model runs faster and gives the best string accuracy when tested on the matched data. Notice that the less number of arcs relates to less computational complexity. The mismatch between AU model and LL_UK data is minimal when compared to AU model tested on US data. This shows that the Australian and UK English accents are more correlated than that of US accented speech. We can clearly see the mismatch in performance between the three different regional accents. Further test results using multiple classifiers and universal hybrid models are tabulated in Table 2 through 4 for LL_UK, LL_AU and LL_US databases.

Table 5 shows the average string accuracy, model size and average arc counts for nine different models. The US system yields the worst and gets about 67.4% and the AU system yields about 75.0% of string accuracy. The UK model is the best among the monolingual system and provides about 76.8% of string accuracy with little more average arc counts than those of US and AU models. MULTI system is better than the AU, UK and US models but inferior to the UNIV system. MULTI_ARC is better than that of individual models, yielding about 88.3% string accuracy, but the arc-count and model size are almost tripled. MULTI_PRN and MULTI_LGL are pretty much providing the same performance and string accuracy of 2% better than that of MULTI_ARC model architecture. This result confirms that by intelligently incorporating the measure between maximum likelihood score and minimum arc-count, one can achieve a a reasonably good accent-independent classifier with improved real-time processing. The MULTI_GRM gives 90.9% string accuracy with less number of arc-counts than those of other multiple classifiers. The MULTI_GRM decoding arc-count is almost halved by restricting a constraint in the grammar during recognition as shown in Fig. 3. Among the four different multiple HMM architectures, the MULTI_GRM configuration yielded the best with less number of arc counts and high string accuracy. We finally observed that the multiple pronunciation for individual words in the lexicon may not be the right choice in accelerating the system robustness due to accent variations.

Overall UNIV_2 outperforms all the other models and yields about 73%, 65%, 62% and 8% in string error rate reductions when compared to the US, AU, UK and MULTI_PRN systems respectively. UNIV_2 exhibits consistent improvements on every LL_AU, LL_UK and LL_US databases. Furthermore, the string accuracies are in the low 90% in all three databases and this suggests that the universal acoustic models can be used for real telephony applications. The average arc count is three times less than the multiple system and comparable with those of the best monolingual systems. Also the model size is twice that of the AU, UK and US models but half the size of multiple system. From the table, it is clear that the UNIV_2 model significantly outperforms the other systems in most cases, as expected. To conclude, the UNIV_2 system provides an efficient way of modeling accent variations from the three languages by using a single Viterbi decoder. It is encouraging that our goal of designing a single global system for all three languages (Australian, American and British English) is achieved by using the universal hybrid system, and the test results have demonstrated the efficacy of enhanced hybrid system. For the sake of comparison, we also built the UNIV_1 model with the same size as that of an individial system. The test results are shown in the eighth column of Tables 2-5. We further observe that the UNIV_1 is still better than that of the individual systems, and comparable in performance with multiple classifiers. If the engineering and economics of power, and size constrained speech processing system in adverse accent environments is the ultimate embedded system, then the UNIV_1 or UNIV_2 universal model methodology is the best choice for successful speech-enabled applications.

5. Conclusions

In this paper, we proposed a framework to address the problem of speech recognition through regional accents for English language. We described several acoustic

modeling techniques to improving the recognizer performance for applications that require mixed-mode accents, and the performance of every system is compared based on various aspects of real-time measures. Universal hybrid modeling system has been proposed and investigated in this paper by intelligently combining data from many different accented speech. The experimental results showed that the universal model in conjunction with suitable model topology to represent the extraneous speech accents not only provide good recognition accuracy but also yield faster response with reduced model size. The major benefit of using an universal hybrid system for a particular language is that there is no need to know about the prior knowledge concerning the nature of the speaker accent that exist in the modern telephone network. In the future experiments, we will apply this universal technique for other languages such as Mandarin (Mainland, Taiwan, HongKong and Singapore chinese accents) [8].

Acknowledgements

The author would like to thank Dr. Rafid Sukkar and Dr. Carl Mitchell for their helpful discussions and support in the early stages of this work.

6. References

[1] L. M. Arslan and J. H. L. Hansen, "Frequency Characteristics of Foreign accented speech", *Proc. ICASSP*, pp. 1123-1126, 1997.

[2] J. R. Bellegarda, "Statistical techniques for robust ASR: Review and perspectives", *Proc. EUROSPEECH*, pp. 33-36, 1997.

[3] R. Chengalvarayan, "A comparative study of hybrid modelling techniques for improved telephone speech recognition", *Proc. ICSLP*, pp. 313-316, 1998.

[4] R. Chengalvarayan, "Improved speech modelling and recognition using a new training algorithm based on outlier-emphasis for non-stationary state HMM", Vol. 26, pp. 191-201, 1998.

[5] R. Chengalvarayan, "Use of multiple classifiers for speech recognition in wireless CDMA network environments", *Proc. ICSLP*, Vol. 3, pp. 386-389, 2000.

[6] R. Chengalvarayan, "Hybrid HMM architectures for robust speech recognition and language identification," *Proc. Systemics, Cybernetics and Informatics*, Vol. 6, pp. 5-8, 2000.

[7] R. Chengalvarayan and L. Deng, "A maximum a posteriori approach to speaker adaptation using the trended hidden Markov model", *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No. 4, pp. 549-557, 2001.

[8] R. Chengalvarayan, "Evaluation of front-end features and noise compensation methods for robusr Mandarin speech recognition", *Proc. EUROSPEECH*, 2001.

[9] V. Fisher, Y. Gao and E. Janke, "Speaker-independent upfront dialect adaptation in a large vocabulary continuous speech recognizer", *Proc. ICSLP*, 1998.

[10] J. L. Hieronymus and S. Kadambe, "Robust spoken language identification using large vocabulary speech recognition", *Proc. ICASSP*, pp. 1111-1114, 1997.

[11] H-G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *International Workshop on Automatic Speech Recognition: Challenges for the Next Millenium*, Paris, France, September 18-20, 2000.

[12] C. H. Ho, S. Vaseghi and A. Chen, "Voice conversion between UK and US accented English", *Proc. EUROSPEECH*, Vol. 5, pp. 2079-2082, 1999.

[13] E. E. Jan, J. B. Ordinas, G. Saon and S. Roukos, "Real-time multilingual HMM training robust to channel variations", *Proc. ICSLP*, pp. 925-928, 2000.

[14] J-C. Junqua, "Impact of the unknown communication channel on automatic speech recognition: A review", *Proc. EUROSPEECH*, pp. 29-32, 1997.

[15] S. Katagiri, B-H. Juang and C-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method", *Proc. IEEE*, Vol. 86, No. 11, pp. 2345-2373, 1998.

[16] J. B. Puel and R. Andre-O'brecht, "Cellular phone speech recognition: Noise compensation versus robust architectures", *Proc. EUROSPEECH*, pp. 1151-1154, 1997.

[17] T. Shultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition", *Proc. ICSLP*, 1998.

[18] J. Sturm and E. Sanders, "Modelling phonetic context using head-body-tail models for connected-digit recognition", *Proc. ICSLP*, pp. 429-432, 2000.

[19] D. L. Thomson and R. Chengalvarayan, "Use of periodicity and jitter as speech recognition features", *Proc. ICASSP*, pp. 21-24, 1998.

[20] O. Viikki, I. Kiss and J. Tian, "Speaker and language independent speech recognition in mobile communication systems", *Proc. ICASSP*, pp. 5-8, 2001.

[21] M. A. Zissman, "Predicting, diagnosing and improving automatic language identification performance", *Proc. EUROSPEECH*, pp. 51-54, 1997.

[22] The European Language Resourece Association web site: *http://www.icpinpg.fr/ELRA/home.html*

[23] http://www.speechdat.org/

Crosslingual Adaptation of Multilingual Triphone Acoustic Models

Andrej Žgank[1], *Bojan Imperl*[1], *Finn Tore Johansen*[2],
Zdravko Kačič[1], *Bogomir Horvat*[1]

[1]Faculty of Electrical Engineering and
Computer Science, University of Maribor,
Smetanova 17, SI-2000 Maribor, Slovenia
andrej.zgank@uni-mb.si

[2]Telenor Research and Development
P.O. Box 83, N-2007 Kjeller, Norway
finn-tore.johansen@telenor.com

Abstract

The paper presents the results of experiments where the influence of adaptation of multilingual triphone speech recogniser in a crosslingual speech recognition was investigated for two different types of multilingual triphone acoustic models. The first type of multilingual triphone acoustic models represents the triphone models generated with the agglomerative clustering technique and the second one represents the triphone models generated with use of the tree-based clustering algorithm. The agglomerative clustering algorithm is based on a definition of 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 is based on the algorithm of Houtgast. The second set of acoustic models was generated using the phonetic decision tree procedure with common set of broad phonetic classes. The crosslingual speech recognition experiments were performed using the multilingual recogniser trained on the Slovenian, Spanish and German 1000 FDB SpeechDat(II) databases for the recognition of the utterances from the Norwegian 1000 FDB SpeechDat(II) database. The mapping of Norwegian phonemes was done with the IPA scheme where five different Norwegian recognition vocabularies were generated. Adaptation of multilingual triphone models was performed on 400 Norwegian utterances. The adaptation significantly improved the recognition results. The best adapted system achieved the recognition accuracy very close to the original – Norwegian – recogniser.

1. Introduction

The speech technology is becoming more and more present in everyday live and is often referred to as one of the key aspects in future development of the technology in various areas. Although the English language is nowadays Lingua Franca, multilingual speech recognition systems are often essential when bringing the speech technology into practice for a wide range of users; for example, when an information retrieval systems are deployed in a multilingual environment. Such systems are expected to support many different languages and it often occurs that for some of the languages there is little or no speech technology resources at all. Since creating the complete speech database for particular language is usually very time consuming and expensive procedure, the crosslingual transfer of speech technology presents the attractive alternative.

This paper describes the experiments that were carried out to investigate the problem of crosslingual transfer of speech technology. It presents the progress of our work in the area of multilingual speech recognition [1, 2, 3]. The goal of the research was to evaluate the advantages of the multilingual speech recognisers based on triphones when porting the recogniser to a new language. The term multilingual speech recogniser denotes the recogniser that was trained using a multilingual speech databases. Previous experiments have indicated [1, 2] that multilingual triphone based recognisers can be efficient when dealing with the unknown languages – languages with little or no speech technology resources.

It is possible to port the existing speech recogniser to a new language without any previous knowledge of the new language. However, some information of the new language is welcome in order to provide the satisfactory recognition accuracy - at least the phonetic dictionary and small amount of additional training sentences [4].

The phonetic dictionary can be generated with grapheme to phoneme rules. There are different ways how to map the phonemes from new language to existing phonemes of multilingual speech recogniser. The mapping can be done following the IPA scheme [5], using a data-driven approach or with a combination of both approaches. Different mapping methods were used in experiments reported by other authors: in [6, 7] the mapping was created with the use of IPA scheme, while the data-driven approach was used in [8]. In our experiments the IPA scheme was applied, due to the specifics of the clustering procedure used.

Previous reports [4, 7] have shown that the performance of the multilingual speech recognition system recognizing a new language significantly improves after the adaptation. This is true even if adaptation of the multilingual speech recognition system was based on a very small amount of speech material of new language. Adaptation of the multilingual speech recognition system to the new language can be performed with the use of many different methods. The simplest ones are with the use of the Baum - Welch parameters estimation [9] or with the Maximum Likelihood Linear Regression [9]. We decided to choose the Baum -Welch algorithm, due to the specifics of the baseline recogniser training procedure.

All our crosslingual experiments are based on the SpeechDat(II) telephone databases [10]. During the experiments two multilingual recognisers based on two different multilingual triphone sets were created. The first multilingual recogniser was based on an agglomerative triphone clustering technique [1] and the second one on a tree-based triphone clustering [11]. For the crosslingual experiments the phonemes from recognition vocabulary of the new language were mapped to existing phonemes of the multilingual models and then the adaptation

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

to the new language was carried out. The crosslingual tests of both unadapted and adapted configurations with mapped recognition vocabulary were performed on both agglomerative and tree-based multilingual speech recognition systems.

The paper is organized as follows. Both clustering procedures and distance measures are presented in Section 2. The databases used are described in Section 3. The experimental systems are presented in Section 4. The crosslingual phoneme mapping and the adaptation technique are described in Section 5 and the results for unadapted and adapted system are presented in Section 6. The conclusion is made in Section 7.

2. Triphone Clustering procedures and distance measures

Two different sets of context dependent multilingual models were used during the crosslingual speech recognition experiments. The first multilingual system is built using the agglomerative (bottom-up) clustering approach [1] and the second one with the phonetic tree-based clustering (top-down) approach [2, 11].

2.1. Agglomerative clustering

The basic idea behind the generation of agglomerative multilingual triphone models proposed in [1] is that the similarity between two different triphones φ_i^L-φ_i+φ_i^R and φ_j^L-φ_j+φ_j^R can be estimated by measuring the similarity of the left phoneme φ^L, central phoneme φ and right phoneme φ^R:

$$\begin{aligned} S_{TRI}(\varphi_i^L\text{-}\varphi_i\text{+}\varphi_i^R, \varphi_j^L\text{-}\varphi_j\text{+}\varphi_j^R) = \\ = W_L\ S(\varphi_i^L, \varphi_j^L) + W_C\ S(\varphi_i, \varphi_j) + \\ + W_R\ S(\varphi_i^R, \varphi_j^R), \end{aligned} \tag{1}$$

where S denotes the similarity between two phonemes, W_L, W_C and W_R are the weights for setting the influence of each phoneme-level similarity estimates, and $S_{TRI}(\varphi_i^L\text{-}\varphi_i\text{+}\varphi_i^R, \varphi_j^L\text{-}\varphi_j\text{+}\varphi_j^R)$ is the resulting similarity estimation of both triphones.

The similarity of two triphones in (1) can be based on any phoneme distance measure. We have decided to apply the phoneme distance measure suggested in [12], which is based on a monophone confusion matrix:

$$\begin{aligned} S(\varphi_i, \varphi_j) &= \frac{1}{2}\sum_{k=1}^{N}[\ c(\varphi_i, \varphi_k) + c(\varphi_j, \varphi_k) \\ &\quad -|c(\varphi_i, \varphi_k) - c(\varphi_j, \varphi_k)|\ ] \\ (\varphi_i, \varphi_j) \in \varphi &\quad , \quad 1 \le i, j \le N\ ,\ i \ne j, \end{aligned} \tag{2}$$

where $S(\varphi_i, \varphi_j)$ denotes the similarity of two phonemes φ_i and φ_j, N is the number of phonemes, $c(\varphi_i, \varphi_k)$ is the number of confusions between phoneme φ_i and phoneme φ_k. The phoneme φ_i should not be the same as φ_j.

When porting the multilingual speech recogniser to new language many unseen triphones may be expected (an example is given in Table 3) as a result of differences in the phonotactics of different languages. During the adaptation to new language the number of unseen triphones can additionally increase since the adaptation dictionary is mapped into phonemes from existing languages. The characteristic of the distance measure defined in equation (2) is, that it can also handle unseen triphones. Each new unseen triphone is compared to all existing triphones using the described distance measure. When the most similar existing triphone with the highest similarity is found, the unseen triphone is tied to this one.

To improve the convergence of the clustering algorithm, a list of polyphones [13] is defined (in our case a list of 14 polyphones was created). These polyphones are used in the clustering procedure. If an average distance among all triphones from the group is less than a predefined threshold T, the group is equated. The average distance between M triphones is calculated as:

$$\begin{aligned} S_M(\hat{\Theta}) &= \frac{\sum_{i=1}^{N}\sum_{j=1}^{N} S_M(\Theta_i, \Theta_j)}{\sum_{i=1}^{N-1} i} \\ (\Theta_i, \Theta_j) &\in \hat{\Theta}\ ,\ 1 \le i, j \le N\ ,\ i \ne j, \end{aligned} \tag{3}$$

where Θ_i denotes the triphone φ_i^L-φ_i+φ_i^R and Θ_j denotes the triphone φ_j^L-φ_j+φ_j^R, $\hat{\Theta}$ is the group of triphones and $S_M(\hat{\Theta})$ is the average distance among all triphones from the group $\hat{\Theta}$.

The result of the clustering algorithm is a list of triphones from all languages that can be tied together. Some triphones that are not similar enough in the sense of equation (3) remain untied.

2.2. Tree-based clustering

The second set of multilingual triphone models that was used in the adaptation experiments is based on a top - down clustering technique. The phonetic decision tree algorithm, as suggested by [11], was used. Each phoneme from all languages used for multilingual speech recognition belongs to different phonetic category (broad class). Similar or equal broad classes from different languages are merged together in one common multilingual broad class. Due to their language specifics, some broad classes are not merged into common category and remain independent. To successfully differentiate between equal phonemes in different languages, each phoneme is tagged with specific language mark. The additional questions [6] about language were not added to the decision tree in our case.

The questions needed for tree building are generated from broad classes. One binary tree is built for each state of the phoneme. The questions are placed in nodes of the tree. At the beginning, all states are placed in the root node of the tree in one cluster. The node is then split into two, by finding the question, which gives the maximum increase of log likelihood for the particular training data set. When the increase is smaller than the threshold, the splitting is stopped.

3. Databases

At the moment, one of the most convenient databases for multilingual speech recognition experiments is the series of SpeechDat [10] databases. These databases provide a realistic multilingual environment for development of voice driven teleservices. Characteristics and recording conditions of all databases were equal, which is crucial in case of crosslingual and multilingual speech recognition experiments. At the moment 25 different SpeechDat(II) databases are available. In our multilingual speech recognition system, the following databases were employed:

- Slovenian (SL) 1000 FDB SpeechDat(II),
- German (DE) 1000 FDB SpeechDat(II),

- Spanish (ES) 1000 FDB SpeechDat(II),
- Norwegian (NO) 1000 FDB SpeechDat(II).

Each database consists of recordings of 1000 speakers over fixed telephone network. Each speaker is represented with approximately 10 minutes of speech. The training part of each database consists of 800 speakers and the remaining 200 speakers were used for test.

Table 1: *The number of training and test utterances, phonemes and size of recognition vocabulary for all used SpeechDat(II) databases.*

Lang.	Train.ut.	Test.ut.	Phon.	Rec.voc.
SL	20658	748	49	605
DE	21463	674	48	674
ES	19164	681	31	681
NO	20346	784	45	792

Due to unsuitableness for training, the following recordings were excluded from the training set of multilingual triphones:

- mispronunciations,
- incomplete utterances,
- unintelligible speech,
- truncated speech.

As seen in Table 1 more than 20.000 sentences from each database were used in training part of experiments described in this paper. In all experiments the W1-W4 corpuses [10], containing phonetically balanced isolated words, were applied during the test. With use of this test corpus, the representation of all phonemes was assured, which is crucial for clustering procedure. Moreover, such tests corpuses do not require use of language models. Test data for each language are also presented in Table 1.

4. Experimental systems

The recognisers applied for crosslingual speech recognition were generated with the use of the script refrec0.9 developed in the framework of the "SpeechDat task force" within COST 249[1] project [14, 15]. During the COST 249 project, different scripts for evaluation of SpeechDat(II) databases in monolingual speech recognition environment were developed. The COST 249 project is continued in the new COST 278 project. The perl script refrec0.9 is created on the base of the HTK toolkit [9] and is an extended version of the tutorial example in the HTK Book.

To achieve more robust performance of the system with telephone speech, a different feature extraction frontend module than in the refrec09 script was applied. The acoustic feature vector consisted of 24 mel-scaled cepstral, 12 Δ - cepstral, 12 $\Delta\Delta$ - cepstral, high pass filtered energy, Δ - energy and $\Delta\Delta$ - energy coefficients. The procedure of maximum likelihood channel adaptation [16] was carried out on feature vectors. The number of elements in the feature vector was reduced to 24 with the use of linear discriminant analysis [16]. With the use of this feature extraction module, the recognition rate was significantly increased.

[1]COST 249 - Continuous Speech Recognition over the Telephone, http://www.elis.rug.ac.be:80/ELISgroups/speech/cost249/

First the monolingual speech recognisers were developed. The 3 state left-right hidden Markov model (HMM) topology was employed for acoustic modeling. The triphone models were built and the number of Gaussian mixtures per state was sequentially increased to 32.

Table 2: *Monolingual recognition results for Slovenian (SL), German (DE), Spanish (ES) and Norwegian (NO) language with monolingual triphone models.*

Language	Recognition rate (%)
SL	88.25
DE	92.51
ES	93.91
NO	78.32

The results of monolingual speech recognition with monolingual systems for all 4 languages are presented in Table 2. The Norwegian monolingual speech recognition system serves as a reference system in our crosslingual adaptation test. The Norwegian monolingual reference system achieved a recognition rate of 78.32%. The COST249 Norwegian system [14] with standard HTK mel cepstral frontend and a bigger vocabulary achieved a recognition rate of 65.27% on the same test set. The Norwegian monolingual reference recognition rate is smaller than monolingual recognition rate for other three languages, as was already noticed in [15, 14]. The probable cause is the length and the number of words in the recognition vocabulary. The size of recognition vocabularies is presented in Table 1.

With the Slovenian, German and Spanish database two multilingual set of triphone models were designed - one for each clustering procedure described in Section 2. The optimal threshold values [2] for both clustering procedures of multilingual triphone models were derived experimentally. With regard to different mapping procedures of Norwegian phonemes to existing phonemes in all three languages (see Section 5.1), the crosslingual speech recognition of Norwegian language without adaptation was first performed. These results, presented in Section 6, served as a reference results for the adaptation experiment. In the next step, the adaptation of multilingual triphone models to Norwegian language on small amount of training material was performed. The test of these adapted models under the same conditions as the unadapted ones was performed.

5. Crosslingual transfer

5.1. Phoneme mapping

For the crosslingual speech recognition the same test set was applied as for the Norwegian reference system. All Norwegian phonemes were mapped to other languages with the use of the IPA scheme. Each Norwegian phoneme was mapped to the IPA symbol of equivalent phoneme in other languages. If the equivalent phoneme did not exist in the target language, the most similar one was chosen (according to IPA notation). The most problematic was the conversion of Norwegian diphthongs. Also problematic was the mapping of Norwegian vowels to Spanish vowels, because the Norwegian language has 18 vowels and Spanish only 5. The mapping resulted in 5 different recognition vocabularies:

- Norwegian to German (ND),
- Norwegian to Spanish (NE),

- Norwegian to Slovenian (NS),
- Norwegian to Multilingual (NM),
- Norwegian to Parallel (NP).

In the case of NM vocabulary Norwegian phonemes were mapped to the optimal phoneme in any of the three target languages. With this procedure the rules of phonotactic were trespassed. In the Norwegian to Parallel (NP) mapping each word in vocabulary has three pronunciation variants: in Slovenian, in German and in Spanish. This way, the vocabulary size was 3 time larger and the data choose the optimal mapping.

As was already found in [3] the crosslingual recognition with Norwegian to Slovenian (NS) mapping without the new reestimation of triphone parameters on Norwegian language, performed worse. Several other mapping possibilities of Norwegian to Slovenian language were tested afterwards, but there was no improvement of the results for NS mapping.

As can be seen in Table 3 for the German language in the case of test, from 2214 triphones 30.67% are missing. As seen in Table 3, the worst case for adaptation and test is the NM vocabulary, where 63.95% and 63.58% of triphones are missing. In the agglomerative system unseen triphones were tied to existing ones with the use of a distance measure and in the tree-based system with the use of a phonetic decision tree. We can see from Table 3, that the German language, which belongs to the same Germanic language group as Norwegian, ought to be the most similar language. The Slovenian, which belongs to Slavic group and the Spanish, which belongs to Romanic group are less similar to Norwegian.

5.2. Adaptation of multilingual triphone models

The adaptation of multilingual triphone models to Norwegian language was performed with few iterations of Baum - Welch algorithm. Only the mean values and the mixture weights of Gaussian distributions in multilingual triphone models were reestimated. There were 400 Norwegian randomly selected sentences from W1-W4 training set used in the adaptation procedure. Collecting such amount of training data does not represent much effort and can be carried out in a minimum time. Only 8.00% of words from the adaptation set overlapped with the same words spoken by different speaker in the test set.

6. Speech recognition results

The first set of tests with Norwegian speech recognition was done on both versions of unadapted multilingual triphone models in crosslingual mode. The test results for all five different configurations of recognition vocabulary mapping with unadapted multilingual triphone models are presented in Table 4. The unadapted tree-based clustering multilingual triphone models outperform the unadapted agglomerative multilingual triphone models. The difference in recognition rate between both versions of unadapted multilingual triphone models is approximately 10%. From all single language mapping vocabularies (ND, NE, NS) that were used for Norwegian crosslingual speech recognition, the German mapping vocabulary performed best in both unadapted systems. With unadapted tree-based triphone models it achieved 43.62% recognition rate and 34.06% with unadapted agglomerative triphone models. This result was anticipated due to experience from mapping.

The mapping of the Norwegian recognition vocabulary to Spanish achieved a recognition rate of 34.57% with unadapted tree-based multilingual models and 19.13% with unadapted agglomerative multilingual models. The lower result for the Spanish mapping was also expected because of the low number of Spanish phonemes, especially the vowels. It is known that vowels are very important for speech recognition. As was already mentioned, several different versions of IPA mapping from Norwegian to Slovenian (NS) were tested in crosslingual recognition without adaptation. No improvement in comparison to results in [3] was achieved. There are two possibilities, which could explain this result. The first and the more probable explanation would be, that Norwegian and Slovenian belongs to too different language groups, to be suitable for crosslingual transfer without adaptation. The second possible explanation is, that IPA charts are not very suitable for such mapping and that some data-driven method of mapping would be more successful.

Table 4: *Recognition rate for crosslingual speech recognition with both unadapted multilingual triphone systems and five vocabulary mapping configurations.*

Map.config	Un. Agglomer. (%)	Un. Tree-based (%)
ND	34.06	43.62
NE	19.13	34.57
NS	1.91	1.28
NM	25.65	32.53
NP	33.42	45.03

In the case of the multilingual vocabulary NM, each triphone can consists of phonemes from different languages. The distance of such unseen agglomerative triphones to existing ones is in average smaller than for other cases of mapping, which shows that similarity between unseen and existing agglomerative multilingual triphones is also smaller. The unadapted NM models achieves recognition rates similar to the unadapted NE system. This is probably due to the trespassing of phonotactic rules. The best mapping configuration in the unadapted tree-based case was when all three mapping possibilities for each word were in the vocabulary (NP). In this case the data choose the best pronunciation variant. The unadapted system achieved a recognition rate of 45.03% with unadapted tree-based multilingual triphone models.

The second set of tests was done on adapted multilingual triphone models, with the same versions of mapped recognition vocabularies as in the case of unadapted models. The results of these tests are presented in Table 5. The performance of all adapted systems was significantly improved.

Table 5: *Recognition rate for adapted crosslingual speech recognition with both multilingual triphone systems and five vocabulary mapping configurations.*

Map.config	Ad. Agglomer. (%)	Ad. Tree-based (%)
ND	48.34	63.52
NE	37.88	55.23
NS	22.96	43.37
NM	27.68	64.80
NP[a]	–.–	–.–

[a] At the time of writing this paper, the tests with adapted multilingual triphone models and NP mapping were not completed yet. However, according to the previous (unadapted) results, very good performance is anticipated.

Table 3: *Number of all mapped triphones and the percentage of missing triphones in tree-based multilingual triphone models for all five adaptation and testing mapping configurations.*

Language	Adapt. Map. Tri.	Adapt. Miss. Tri.	Test Map. Tri.	Test Miss. Tri. (%)
ND	1417	28.93	2214	30.67
NE	1130	29.82	1606	32.50
NS	1349	29.50	2085	44.41
NM	1473	63.95	2331	63.58
NP	3896	29.39	5905	36.02

As can be seen from Tables 4 and 5, the performance of all systems after the adaptation has significantly improved. With and without the adaptation, the tree-based triphone models outperform the agglomerative triphone models. In the case of adapted agglomerative triphone models, the best mapping configuration is still the ND. The recognition rate increased from 34.06% to 48.34%. When NE mapping was used with adapted agglomerative models, the results improved from 19.13% to 37.88%. The improvement for NS adapted agglomerative models is from 1.91% to 22.96%. The smallest improvement achieved was in the case of NM mapping, from 25.65% to 27.68%. Additional experiments will be performed in the future to clarify the small improvement of the NM mapping.

It is worth of noting that after the adaptation, the adapted tree-based triphone models with NM mapping (increase from 32.53% to 64.80%) outperformed the ND mapping (increase from 43.62% to 63.52%), despite the trespassing of phonotactic rules in NM mapping. Also a large increase in recognition rate from 1.28% to 43.37 % for NS mapping was achieved. The result probably confirms the hypothesis, that Norwegian and Slovenian language are too dissimilar, to be used in crosslingual speech recognition transfer without the adaptation. The best adapted system (Table 5) with tree based multilingual triphone models and NM mapping vocabulary achieved the recognition rate of 64.80%[2], which is very close to the performance of the reference Norwegian system (Table 2).

The Table 6 contains the analysis of results for Norwegian to Parallel mapping for unadapted version of multilingual triphone models. The percentage of selected language pronunciation variants from the parallel recognition vocabulary was calculated. Then the percentage was calculated, how many of these selected variants per language were correct. The percentage of selected variants for agglomerative and tree-based multilingual triphone models are almost identical, but the percentage of correct variants is lower in the case of the agglomerative multilingual triphone models. As can be seen, the highest part of the selected variants, with the highest correct rate, belongs to German pronunciation variant for both cases of multilingual triphone models. The recognition rate of unadapted tree based multilingual models with NS mapping was only 1.28%, but in the case of NP mapping with the same models, 27,27% of selected Slovenian pronunciation variants were correct for tree-based multilingual triphone models. But it must be also taken into account, that the number of selected Slovenian variants is very small, which supports the hypothesis that Norwegian and Slovenian language are very dissimilar.

[2]Completion of the tests with adapted multilingual triphone models and NP mapping is expected to yield even higher recognition rates.

Table 6: *Analysis of NP recognition results for both versions of multilingual triphone models.*

	Agg. (%)		TB (%)	
Lang.	select	correct	select	correct
DE	71.81	36.94	71.68	48.75
ES	26.91	25.12	26.91	36.02
SL	1.28	10.00	1.41	27.27

7. Conclusion

In the paper the crosslingual adaptation of agglomerative and tree based triphone models to Norwegian language was investigated. The best adapted system with tree based multilingual triphone models and NM recognition vocabulary came very close to the original Norwegian reference system. This indicates that crosslingual transfer of speech technology based on multilingual triphone modeling might be a promising approach when dealing with unknown languages.

However, the best type of multilingual triphone modeling still needs to be determined. So far the tree-based multilingual triphone models were found to be the most efficient for porting a recogniser to a new languages. The agglomerative multilingual models that performed best in the case of multilingual speech recognition experiments [2] did not prove that well in the case of crosslingual experiments – exact reasons for this still needs to be determined.

All crosslingual experiments were based on the phoneme mapping with IPA charts. Such mapping did not perform well in the case of monophone multilingual models, but has performed very good in the case of the multilingual triphone models. Perhaps this is the result of the NP mapping, where unlike the previous experiments with the monophone models, various mapping hypotheses were suggested and the data was left to choose the optimal mapping. The experiments have yielded promising results but is to early to draw any general conclusions. Only four languages were used during the described tests therefore the scale of experiments (the number of languages) needs to be augmented. Next, other mapping techniques for unadapted crosslingual transfer should be investigated. As indicated during the experiments, the results are likely to depend on the language groups that the languages involved in the experiment belongs to. In future, experiments should be also carried out for the languages within the same language group and than extended across all language groups to eventually build the ”global” set of triphone models.

Acknowledgments

The authors wish to acknowledge Siemens AG and the Universitat Politechnica de Catalunya for providing the Ger-

man and Spanish SpeechDat databases. The authors also wish to thank Melita Zemljak (Faculty of Education, University of Maribor) for her help with the IPA scheme.

8. References

[1] Imperl, B. and Horvat, B., "The Clustering Algorithm for the Definition of Multilingual Set of Context Dependent Speech Models", Proc. EUROSPEECH'99, Budapest, Hungary, 1999.

[2] Imperl, B., Kačič, Z., Horvat, B., Žgank, A., "Agglomerative vs. Tree-Based Clustering for the Definition of Multilingual Set of Triphones", Proc. ICASSP'2000, Istanbul, Turkey, 2000.

[3] Žgank, A., Imperl, B., Johansen, F. T., Kačič, Z., Horvat, B., "Crosslingual Speech Recognition with Multilingual Acoustic Models Based on Agglomerative and Tree-Based Triphone Clustering", Proc. Eurospeech 2001, Aalborg, Denmark, 2001.

[4] Bub U., Köhler, J., Imperl, B., "In-Service Adaptation of Multilingual Hidden-Markov-Models", Proc. ICASSP'97, Munich, Germany, 1997.

[5] "The IPA 1989 Kiel Convention", Journal of the International Phonetic Association 1989(19) pages 67 – 82.

[6] Schultz, T. and Waibel, A., "Multilingual and Crosslingual Speech Recognition", Proc. DARPA Broadcast News Workshop 1998, Lansdowne, USA, 1998.

[7] Schultz T. and Waibel A., "Polyphone Decision Tree Specialization for Language Adaptation", Proc. ICASSP'2000, Istanbul, Turkey, 2000.

[8] Köhler, J., "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds", Proc. ICSLP'96, Philadelphia, USA, 1996.

[9] Young, S., "The HTK Book", Cambridge University.

[10] Höge, H., Tropf, H., Winski, R., van den Heuvel, H., and Haeb-Umbach, R., "European Speech Databases for Telephone Applications", Proc. ICASSP'97, pages 1771 – 1774, Munich, Germany, 1997.

[11] Young, S., Odell, J., Woodland, P., "Tree-based State Tying for High Accuracy Acoustic Modelling", Proc. ARPA Human Language Technology Conference, Plainsboro, USA, 1994.

[12] Andersen, O., Dalsgaard, P., and Barry, W., "Data-driven Identification of Poly- and Mono-phonemes for Four European languages", Proc. EUROSPEECH'93, pages 759 – 762, Berlin, Germany, 1993.

[13] Imperl, B. "Clustering of context dependent speech units for multilingual speech recognition", Multi-lingual Interoperability in Speech Technology, Proc. ESCA-NATO Tutorial and Research Workshop, Leusden, Netherlands, 1999.

[14] Johansen, F. T., Warakagoda, N., Lindberg, B., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G., "The COST 249 SpeechDat Multilingual Reference Recogniser", XLDB - Very Large Telephone Speech Databases, LREC'2000 Workshop Proc., Athens, Greece, 2000.

[15] Lindberg, B., Johansen, F. T., Warakagoda, N., Lehtinen, G., Kačič, Z., Žgank, A., Elenius, K., and Salvi, G., "A Noise Robust Multilingual Reference Recogniser Based on SpeechDat(II)", Proc. ICSLP'2000, Beijing, China, 2000.

[16] Haunstein, A. and Marschall, E., "Methods for Improved Speech Recognition Over the Telephone Lines", Proc. ICASSP'95, Detroit, USA, 1995.

Multilingual Text-To-Phoneme Mapping for Speaker Independent Name Dialing in Mobile Terminals

Kåre Jean Jensen, Søren Kamaric Riis, and Morten With Pedersen

Nokia Mobile Phones, Copenhagen, Denmark
{kaare.jensen, soren.riis, morten.w.pedersen}@nokia.com

Abstract

This paper presents recent experiments on multilingual Text-To-Phoneme (ML-TTP) mapping for speaker independent name dialing using the low complexity ML-TTP method first introduced at EuroSpeech 2001. The ML-TTP method avoids the use of a Language IDentification (LID) module and produces phoneme transcriptions for different languages using the same model. Through an extensive comparison on an 11 language name dialing task it is illustrated that the ML-TTP method yields a similar performance to that of a baseline LID/TTP method, but at a significant saving in model size and real-time decoding complexity.

1. Introduction

Speaker independent speech recognition systems based on multilingual phoneme models have recently attracted significant interest [6, 7, 10]. These systems are designed to handle several different languages simultaneously and are based on the observation that many phonemes are shared among different languages [6, 7, 10]. The basic idea in multilingual acoustic modeling is to estimate the parameters of a particular phoneme model using speech data from a number of different languages that include this phoneme. Sharing of phoneme models across languages can significantly reduce the number of free parameters in the system and thereby the memory requirements compared to using separate phoneme models for each language. It has been observed that for many languages sharing a common phoneme set, good performance can be obtained using a multilingual recognizer designed from a only few representative languages. This alleviates the need for large acoustic speech databases for all languages to be supported. Such multilingual recognizers are very attractive for e.g. name dialing and command word recognition applications on world wide portable products with limited computational resources.

In speaker independent name dialing applications for mobile terminals the user is allowed to change the active vocabulary online. That is, the user can add, delete, or modify names in the phonebook located on the device, thereby changing the active vocabulary. For such applications, a method for transcribing the written words into phonetic transcriptions on-the-fly is needed. For a monolingual system this is fairly straightforward, by using a language dependent Text-To-Phoneme (TTP) model for mapping grapheme (letter) sequences into phoneme sequences. Typical approaches for TTP mapping include dictionary lookup tables, rules, and statistical mapping methods like decision trees [4] and neural networks [5, 9]. For multi-lingual speech recognition applications, the task of transcribing a vocabulary word using a language dependent TTP model is complicated by the fact that the language of the vocabulary words may not be known a priori. Thus, explicit information about the language of the word must be available. If this information is not available, a statistical method for Language IDentification from text (LID) can be applied. Viable approaches for language identification are decision trees, N-grams [3] and neural networks. These methods are all very accurate if a sufficient amount of text is available. However, for applications like name dialing, it may be very difficult to assign a language tag to a particular name as the amount of text is very limited. In fact, a unique language tag may not always exist as a name may have the same orthographic form in several languages even though it is pronounced quite differently.

In [7], a novel low complexity approach denoted multilingual TTP (ML-TTP) mapping for generating pronunciations from written text on-the-fly was introduced for a system comprising four different languages. The main advantage of the ML-TTP approach is that no LID module is needed as a single TTP model handles the phoneme mappings for all languages. As similar mappings in different languages are only represented once in the ML-TTP module it will typically also be somewhat less memory consuming than the approach based on a LID module and multiple language dependent TTP models. For portable devices with limited computational and memory resources, a reduced memory requirement is particularly attractive.

In this paper we report the results of an extensive comparison between the ML-TTP method and the method based on LID and language dependent TTP. The comparison is conducted on a speaker independent name dialing task using a multilingual recognizer supporting 11 different European languages.

2. Language Dependent TTP Models

As mentioned above, the TTP mappings in a multilingual system can be handled by a combination of a LID module and a set of language dependent TTP models. This section describes the LID and TTP modules used in this work.

2.1. Language Identification

The LID module is a statistical model based on a neural network classifier. In a set of initial experiments the neural network classifier was found to yield a superior LID classification performance compared to both decision trees and N-grams. Furthermore, the neural network LID model makes a good compromise between complexity and performance. The network architecture is a standard multi-layer perceptron (MLP) with softmax normalized outputs [1]. Each output unit (class) corresponds to a language and with the softmax normalization the outputs will approximate class posterior probabilities.

The language of a given word or word sequence is estimated by processing one letter at a time and then combining

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

the results, producing a list of the most probable languages. In order to include letter context within the word, the input to the network is a window of letters with the focus letter placed in the center. This window is then slid across the word, placing each letter of the word in the center of the window. Each letter of the input window is encoded using an orthogonal binary vector in order to avoid false correlations between letters, see [5, 9] for details. The LID classifier gives an estimate of the posterior probability that the focus letter belongs to each of the languages supported by the model. However, for LID classification we need a score indicating to which language the entire word belongs. One method, the *voting method*, is to simply find the most probable language at each window position and then select the language that wins for most letters. Another method is based on a confidence score, which for a word w with orthographic form $\{l_1, l_2, \ldots, l_L\}$ is defined by:

$$A_i(w) = \frac{a_i(w)}{\sum_{k=1}^{C} a_k(w)}, \quad a_i(w) = \prod_{j=1}^{L} O_i(l_j) \qquad (1)$$

In Equation 1, A_i is the confidence score for language i, C is the number of languages modeled by the LID network, and $O_i(l_j)$ is the i^{th} output of the neural network when letter l_j is placed in the center of the input window. The winning language is then simply selected as the one which gives the largest confidence score. The confidence score can be used to create an N-best list by ranking the languages according to their confidence scores. In a set of initial experiments this method was found to provide better classification accuracy than the voting method described above.

When a LID N-best list is used together with language dependent TTP models, one transcription is simply included for each language in the N-best list. This increases the vocabulary size and thereby the real-time decoding complexity, but it also improves the recognition performance. The increase in decoding complexity can be reduced somewhat by organizing the phoneme transcriptions in a tree structure where similar initial phoneme transcriptions are only decoded once. Furthermore, the confidence score in Equation 1 can be used to reduce the N-best list by only including transcriptions for those languages in the N-best list that has a confidence score larger than some predefined threshold.

2.2. Pronunciation Rules

For many languages the TTP mapping can be performed with high accuracy using a set of pronunciation rules. For such languages, the pronunciation rules used in this paper define simple textual substitutions between grapheme and phoneme segments. If more than one rule is applicable the mapping corresponding to the rule with the longest grapheme segment is used. The rule sets were found to be very compact and accurate for many languages including e.g. Finnish and Italian.

For some languages like e.g. Spanish, the rule sets grow quite large so even though they are accurate, a neural network TTP model was found to be more compact while still providing a similar mapping accuracy.

2.3. Neural Network Based TTP Models

For non-rule based languages like e.g. English and German a standard MLP was used. The neural network TTP model has been found to provide a good compromise between complexity and mapping accuracy [5].

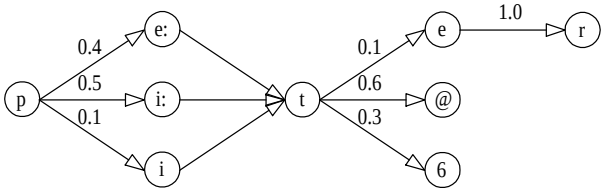


Figure 1: Pronunciation of name “Peter” in German (p-e:-t-6), English (p-i:-t-@) and Spanish (p-i-t-e-r) arranged as a branched grammar. The values on the arcs between phonemes indicate probabilities of the phonemes as provided by e.g. the TTP module. SAMPA notation is used for phonemes.

Like the LID module, the language dependent neural network TTP model is a standard MLP with softmax normalized outputs. For the TTP model, each output unit (class) corresponds to a phoneme. The input data for the model is similar to that of the LID model: a letter window is slid across the word producing an output vector for each letter in the word. That is, for each letter we get a probability estimate that this letter (and context) corresponds to each of the phonemes defined in the model. The phonetic transcription of the word is obtained by selecting the phoneme having the highest probability for each letter.

A neural network model produces exactly one output (vector) for each input (vector). Thus, for neural network TTP models the number of phonemes must equal the number of letters in the corresponding word. However, in many cases the phonetic transcription is shorter than the word. This problem is solved by introducing an “empty” phoneme denoted a *null phoneme*. The null phonemes are inserted into the training data by an alignment algorithm as described in [2]. When language dependent transcriptions are generated by the neural network model all null phonemes are simply removed.

In a few cases the phoneme transcriptions are longer than the corresponding words. In these cases the alignment is done by the use of *pseudo phonemes*. A pseudo phoneme is a concatenation of two phonemes which is then regarded as an individual phoneme and added to the phoneme set. When language dependent transcriptions are generated the pseudo phonemes are expanded to their corresponding regular phonemes.

3. Multilingual TTP Mapping

The concept of a *branched grammar* in combination with TTP mapping was first introduced in [7] as a way of avoiding the LID module and instead generating a single multilingual phonetic transcription. The idea of the branched grammar method is to allow the transcription of a word to contain more than one possible phoneme at any given position. Figure 1 shows a branched grammar transcription for the name Peter for the three languages English, German, and Spanish. The two letters ‘p’ and ‘t’ transcribe into the ‘p’ and ‘t’ phonemes for all three languages. The vowels, on the contrary, have three different possibilities and therefore three branches. This way we have three different pronunciations incorporated into one single transcription. If the ML-TTP model provides phoneme probability estimates, a score can be assigned to each of the different branches. This score can be used as a transition weight during decoding.

When the branched transcription is decoded for a given ut-

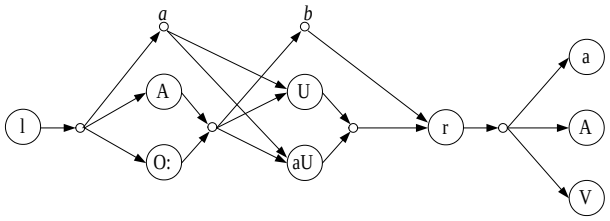


Figure 2: Recognition time model for the word “Laura” showing how the acoustic phoneme models are connected in the case of a branched grammar with null phonemes in the branches. SAMPA notation is used for the phonemes.

terance, a branch with the best match to the utterance will give the highest contribution to the decoding score. If the name “Peter” is pronounced in German for the example in Figure 1, the upper branch for phoneme position two and the lower branch for phoneme position four will make the major contribution to the final decoding score.

Probability estimates produced by the ML-TTP model can also be used to select the number of phonemes for a branch. This is done by introducing a probability threshold so that only phonemes with a probability larger than the threshold will be included in a branch. If no phonemes have a probability higher than the threshold only the most probable one is included.

In this work the ML-TTP modeling is handled by a standard MLP neural network similar to the language dependent TTP networks. The only differences are that a multilingual phoneme set is used as output classes, and that the network is trained using transcriptions from several languages.

For the non-branched case, like e.g. the language specific TTP, null phonemes in the output transcription are deleted. However, in the branched case the null phonemes are a bit more complicated to handle. If we have a null phoneme in a branch, the recognition time model will have paths going into the branch but also paths completely skipping the branch. If there are several successive branches with null phonemes, this potentially means that a large number of phonemes in the transcription can be skipped. This may increase confusion between entries in the vocabulary causing the recognition rate to drop significantly. In other words, the temporal constraints which are needed to discriminate the entries of the vocabulary may become too relaxed.

Several approaches to alleviate the null phoneme problem in the branches have been attempted with limited success: e.g. to simply delete the null phonemes, to delete the null phonemes and phonemes having a lower probability, to skip the whole branch if the null phoneme has the highest probability, etc. The best performance was achieved by including all null phonemes in the branches and forcing the paths in the recognition time model to skip no more than one phoneme position. This is illustrated in Figure 2 where the recognition time model for the word “Laura” is shown. The arrows represent all different paths through the model and the small circles are non-emitting *glue states*. The two glue states, *a* and *b*, represent null phonemes in the two branches. From the initial phoneme ‘l’ three paths go to phoneme ‘O:’, ‘A’, and glue state *a*. From glue state *a*, two paths go to phoneme ‘U’ and ‘aU’ but no path to the null phoneme in the next branch represented by glue state *b*. By excluding the path from glue state *a* to *b* we force the path through glue state *a* to only skip the first branch. That is, no paths through the model can skip *both* the first branch [A, O:] and the second branch [U, aU].

Language	N_{utt}	Language	N_{utt}
Czech	3 950	Italian	5 754
Dutch	5 034	Polish	4 614
English	5 038	Spanish	5 283
Finnish	3 600	Swedish	7 440
German	7 979	Turkish	2 922
Hungarian	7 470	*All*	*59 084*

Table 1: Name dialing datasets for 11 European languages used for evaluating multilingual recognizer system performance for various TTP methods. The test set for each language is based on a 120 word vocabulary of names (90 full names, 10 given names and 20 foreign names). N_{utt} is the number of test utterances

4. Experimental Setup

In this work the ML-TTP method was compared to a method employing a LID module in combination with language dependent TTP modules. The comparison was conducted on a speaker independent name dialing task on the 11 different European languages listed in Table 1. A multilingual mono-phoneme based recognizer supporting all 11 languages simultaneously was used for all experiments.

For testing the overall system recognition rate a Nokia in-house test set for each of the 11 languages listed in Table 1 was used. The test set for each language is based on a 120 word vocabulary of names (90 full names, 10 given names and 20 foreign names).

Details about architecture and training of the multilingual recognizer, the TTP and the LID models are given below.

4.1. Multilingual acoustic module

The acoustic models in the multilingual recognizer were designed using speech data from four languages only: Finnish, German, English (US and UK), and Spanish. For these four languages the total number of mono-phonemes is 133 corresponding to 39 phonemes for English, 28 for Spanish, 43 for German and 23 for Finnish. By defining a common multilingual phoneme set sharing similar phonemes, the total number of mono-phonemes can be reduced to 67 without affecting overall system performance, see [10] for details.

Even though the acoustic models have been trained using data from the above mentioned four languages only, they have been observed to give good performance when used for other languages based on phonemes contained in the 67 multilingual phoneme set. Thus, as shown in Section 5 below, good performance is obtained on all 11 European languages using this set of 67 multilingual acoustic models.

The acoustic phoneme models in this work were based on a low complexity hybrid known as Hidden Neural Networks (HNN) [7, 8]. The multilingual HNN acoustic models take up a total of 17 kb of memory. However, the TTP methods discussed in this paper apply equally well to an HMM based multilingual phoneme based recognizer. Please refer to [7] for further details concerning preprocessing, architecture and training of the hybrid multilingual recognizer.

During decoding of the HNN, an all-path, tree-structured forward decoder was applied as this has been observed to yield a better performance than Viterbi decoding for HNN acoustic models [8]. By using a forward decoder all paths in the branched grammars produced by the ML-TTP module can contribute to the overall score of a particular vocabulary entry. On the other hand, when a Viterbi decoder is used, only a single

path through the branched grammar will contribute to the overall score of the vocabulary entry.

4.2. TTP mapping module

The number of inputs to the different TTP networks used for non-rule based languages and the ML-TTP approach were determined by the input context size and the number of different letters in the language. All neural network TTP models, except the one for Spanish, used a context of four letters on both sides of the central letter in the input window. For the Spanish neural network TTP model, a context of only one letter on both sides of the central letter was found to provide very accurate TTP mappings. The model sizes (using 8 bit/parameter precision) of the various neural network TTP models are listed in Table 2.

The neural network TTP models were trained by stochastic online back-propagation using lexicons of phonetically transcribed words. All training material for the language dependent neural network TTP models was taken from the CMU-Dic (English), LDC-CallHome-German and LDC-CallHome-Spanish pronunciation lexicons. The number of words extracted from these lexicons for training is shown in Table 2.

To obtain a training set for the ML-TTP module it was necessary to do phonetic transcription of a set of words for all the rule-based languages. For this purpose, a word list was generated from various Internet textual resources for each of the rule-based languages. These word lists were then processed using the TTP rules to produce training data of phonetically transcribed words. Table 2 lists the number of *different* words available for each of the rule-based languages for ML-TTP training. As seen from the table, the number of words for some of the languages is very limited compared to e.g. the English or German databases. To ensure that the ML-TTP training is not dominated by the English and German examples, a balanced training set was created by using roughly 50 000 examples for each of the languages. For English and German, 50 000 examples where chosen at random from the large databases also used for training the language dependent neural network TTP models. For all other languages, the available datasets where repeated as many times as needed in order to generate roughly 50 000 examples for each language. Naturally, the low number of *different* words for e.g. Czech and Polish may pose a problem, as the distribution of phonemes in the generated training data may differ significantly from the distribution of phoneme "normally observed" in these languages. Similarly, the limited number of words implies that the grapheme segments defining the mapping rules only occur in a very limited number of orthographic contexts. This may hamper learning of the mappings in the rule-based languages by the ML-TTP network. A larger and more representative set of different words can alleviate such effects.

For the rule-based languages Table 2 lists the sizes of the rule set files.

4.3. LID module

If the recognition system is required to simultaneously support a fairly small number of languages, like e.g. 2-4, a pronunciation for each supported language could be generated for each word instead of using a LID module. However, if the system is required to support a large number of languages as considered in this work, this approach may be computationally prohibitive for embedded devices, as the size of the active vocabulary is "artificially" increased. Therefore, a standard MLP was used for language identification in this work. The LID network had

Language	Model	Model size	Words available
English	Network	30 kb	87 215
German	Network	10 kb	255 327
Spanish	Network	0.7 kb	36 494
Czech	Rules	0.6 kb	855
Dutch	Rules	0.7 kb	7 346
Finnish	Rules	0.2 kb	15 132
Hungarian	Rules	1.1 kb	3 640
Italian	Rules	1.3 kb	11 502
Polish	Rules	1.6 kb	915
Swedish	Rules	0.9 kb	5 437
Turkish	Rules	0.3 kb	2 056
ML-TTP	Network	32 kb	583 413

Table 2: TTP methods for the different languages. The column denoted "Words" shows the number of *different* transcribed words available for each language (except for the ML-TTP case, see text).

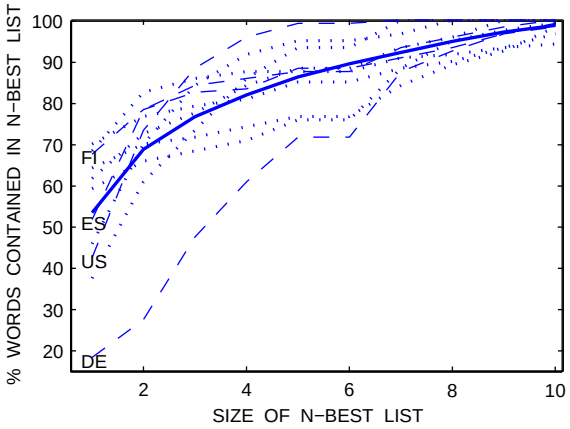


Figure 3: LID language classification performance on the test vocabulary for the 11 languages as a function of the N-best list size.

one output for each of the 11 languages and a context of 4 letters on each side of the central letter in the input window. With 8 bit/parameter precision the size of the LID network was 22 kb. Thus, the overall memory consumption of the LID and language dependent TTP models was about 70 kb or more than 2 times the size of the ML-TTP network.

As the ML-TTP network, the LID network was trained using stochastic online back-propagation on a balanced training set containing roughly 50 000 examples for each language. The dataset used for training the LID network was obtained by simply replacing all phoneme transcriptions by language tags in the ML-TTP training set.

5. Results

This section describes the results that were obtained on the test sets for the 11 languages. First, the results using a LID module in combination with language dependent TTP modules are described. Then follows the results using an ML-TTP module, and, finally, the two approaches are compared.

When using a LID module in combination with language dependent TTP modules one has to decide how many languages to include in the LID module N-best list, i.e. how many tran-

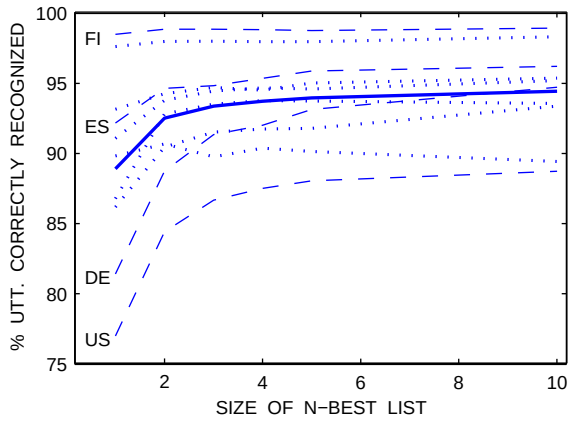


Figure 4: Recognition performance using LID and language specific TTP models as a function of the N-best list size.

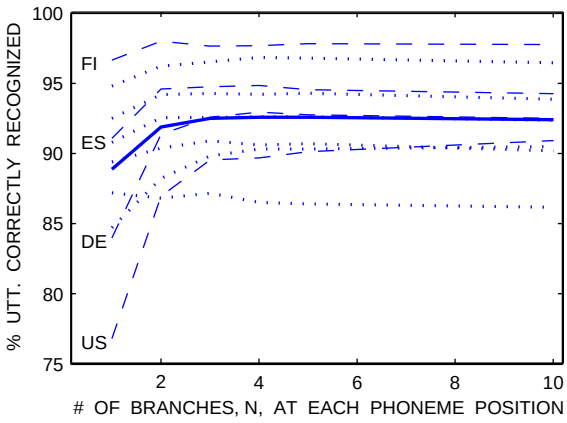


Figure 5: Recognition performance using a 32 k ML-TTP model as a function of the number of branches.

scriptions to include of a particular word in the vocabulary. Naturally, the performance of a recognizer when using this method depends on whether or not the correct language for the word is included in the N-best list; the larger the N-best list, the more likely it is included. Figure 3 displays the success rate of the LID module in this respect, as measured on the test set vocabulary for each of the 11 languages. The thick, solid line shows the average percentage over the 11 languages of the test words where the correct language ID was contained in the N-best list. The rest of the lines show the percentage for each language individually. The results for the languages on which the acoustic models were trained (Finnish, German, Spanish, and English) are shown with dashed lines. From Figure 3, a large difference in the LID classification rate is observed for the different languages. Part of the difference can be attributed to the character sets of the different languages. As the letters 'a' to 'z' are used in all 11 languages, those words that contain special language specific characters are easier to classify correctly than words only containing characters 'a' to 'z'. The poor classification rate for German and English for small N-best lists is due to the fact, that the test set vocabularies for these two languages contain words mainly using letters from 'a' to 'z'. The low average LID classification rate for small N-best lists indicates the difficulty of language identification from very short text segments.

Figure 4 displays the performance of the recognizer as the size of the LID N-best list is increased. The thick, solid line shows the average recognition accuracy as measured across all 11 languages, and the rest of the lines show the recognition accuracy for each language individually. Note how the accuracy increases significantly up until an N-best list of size 3, after which there is only a modest gain in performance by increasing the size of the N-best list. Despite the large increase in the LID module success rate beyond an N-best list of size 3, as seen in Figure 3, limitations of the language dependent TTP models and of the acoustic models apparently limit the increase of the recognizer performance. Hence, in this case an N-best list of size 3 provides a good trade-off between performance and complexity.

We now investigate the use of the ML-TTP module. When using an ML-TTP module one needs to determine how many alternative phonemes to include at each phoneme 'position', i.e. how many branches to allow at each position. Figure 5 displays the performance of the recognizer as the number of

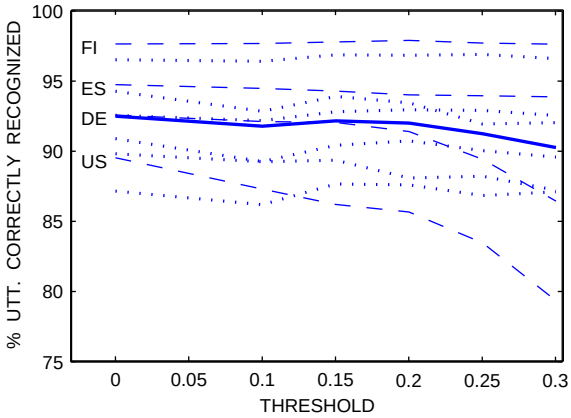


Figure 6: Recognition performance using a 32 k ML-TTP model with a maximum of 3 branches as a function of the threshold.

phoneme branches is increased when using an ML-TTP model with 32 k parameters. It is seen that the performance increases significantly by introducing the branching. However, beyond a branching factor of $N = 3$ there is hardly any improvement in the performance. This is explained by the fact that the additional phoneme alternatives beyond $N = 3$ have a very small probability and do not contribute to the modeling of the acoustic signals.

In order to reduce the complexity of the decoder when using a branched grammar, one may introduce a phoneme probability threshold, as described in Section 3. Phonemes with a probability below the threshold are excluded from the transcription. Figure 6 displays the performance of the recognizer as the threshold for the ML-TTP module output is increased. It is seen that the recognizer performance remains largely unchanged up to a threshold of about 0.2, after which it starts to deteriorate, as important phonemes are now being excluded from the transcriptions.

Figure 7 compares the relative decoding complexity in terms of number of states when using lexicons generated by different TTP methods. The complexity is measured in terms of the number of states in the recognition time model constructed from a specific lexicon. The relative complexity is obtained by normal-

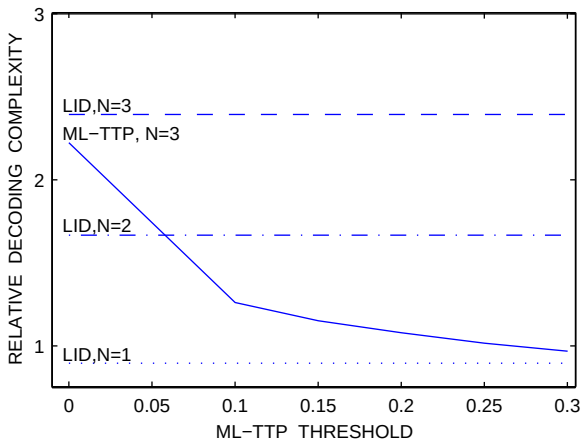


Figure 7: Comparison of relative complexity of LID and ML-TTP as a function of ML-TTP threshold. A maximum of 3 branches was used in the ML-TTP transcriptions.

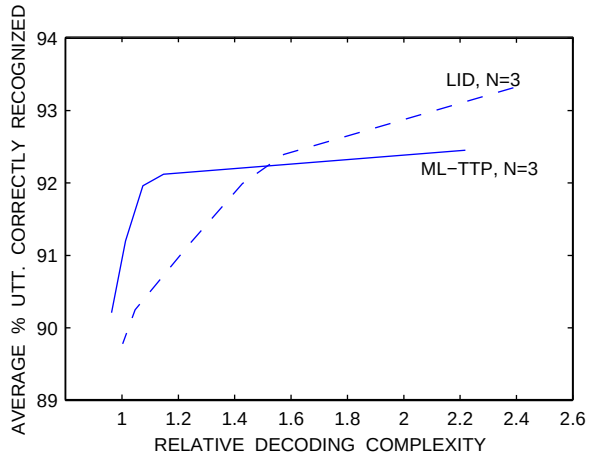


Figure 8: Performance versus decoding complexity for LID and ML-TTP. An N-best list of size 3 was used in the LID module and a maximum of 3 branches was allowed in the ML-TTP model. The decoding complexity was controlled by varying the LID and ML-TTP thresholds.

izing with the number of states resulting from a lexicon generated using language dependent TTP models and know language ID. From the figure, it is seen that by increasing the size of the LID N-best list the decoding complexity is increased significantly. On the other hand, by using an ML-TTP model with a maximum of 3 branches and a threshold of 0.2 the decoding complexity is only increased by 10 %, while maintaining a good recognition performance as seen in Figure 6.

When creating transcriptions based on a LID module N-best list, one may omit transcriptions for languages having a score below some threshold, and thus considered less likely. This will naturally decrease the decoding complexity of the resulting recognizer, however, at the price of a reduced recognition performance. Figure 8 displays the relative complexity and the performance of the LID based method, using an N-best list of size 3, as the score threshold is varied. For the LID based method, it is seen that the recognition performance decreases rapidly with decreasing complexity. Figure 8 also displays a similar curve for the ML-TTP based method. Contrary to the LID based method, the performance of the ML-TTP method is basically not affected by decreasing the complexity until a certain point at which performance falls of rapidly. It is seen that the LID based recognizer reaches the best overall performance, however at the cost of a large decoding complexity. In devices where computational complexity and memory consumption are important issues, the ML-TTP based recognizer provides an attractive solution with competitive performance.

6. Conclusion

A novel approach denoted ML-TTP for generating multilingual text-to-phoneme mappings have been evaluated on a large name dialing task for 11 European languages.

Compared to a system employing a language identification module and language specific TTP models, the ML-TTP approach was found to provide competitive performance, but at a significant saving in decoding complexity and memory usage.

If the LID module provides an N-best list of alternative languages, it is possible to generate several transcriptions for each vocabulary entry. For sufficiently large N-best lists it was found that the method based on a LID module and language dependent TTP models can give slightly better performance than the ML-TTP method. However, in this case the decoding complexity of the LID based method is at least doubled compared to the ML-TTP system.

7. References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[2] T. Dutoit. *An Introduction to Text-to-Speech Synthesis.* Kluwer Academic Publishers, 1997.

[3] G. Grefenstette. Comparing Two Language Identification Schemes. In *Proceedings of 3rd International Conference on Statistical Analysis of Textual Data*, pages 1–6, 1995.

[4] J. Häkkinen and J. Tian. Decision Tree based Text-To-Phoneme Mapping for Speech Recognition. In *Proceedings of ICSLP*, 2000.

[5] K. J. Jensen and S. Riis. Self-Organizing Letter Code-Book for Text-To-Phoneme Neural Network Model. In *Proceedings of ICSLP*, 2000.

[6] F. Palou, P. Bravetti, O. Emem, V. Fischer, and E. Janke. Towards a Common Phone Alphabet for Multilingual Speech Recognition. In *Proceedings of ICSLP*, pages 1–1, 2000.

[7] S. Riis, M. W. Pedersen, and K. J. Jensen. Multilingual Text-To-Phoneme Mapping. In *Proceedings of EuroSpeech*, 2001.

[8] S. Riis and O. Viikki. Low Complexity Speaker Independent Command Word Recognition in Car Environments. In *Proceedings of ICASSP*, 2000.

[9] T. J. Sejnowski and C. R. Rosenberg. Parallel Networks that Learn to Pronounce English Text. *Complex Systems 1*, pages 145–168, 1987.

[10] O. Viikki, I. Kiss, and J. Tian. Speaker- and Language-independent Speech Recognition in Mobile Communication Systems. In *Proceedings of ICASSP*, 2001.

Multilingual Speech Interpretation for Cellular Phones

Yasunari Obuchi, Yoshinori Kitahara, and Atsuko Koizumi

Central Research Laboratory
Hitachi, Ltd.
Kokubunji, Tokyo 185-8601, Japan
e-mail: obuchi@crl.hitachi.co.jp

Abstract

This paper describes the development and field trial tests of the multilingual speech interpretation system. The system is based on the fixed sentence interpretation, in order to achieve a high recognition rate and to guarantee the translation accuracy. More than 1,700 sentences are registered, and a simple network grammar enables the system to deal with various expression forms for the same meaning. The first field trial was carried out with the Japanese-to-Korean interpretation system. We received 6,753 calls during two months, and the data showed that the biggest problem was the out-of-vocabulary (OOV) inputs, which resulted in 39% task completion rate. The second prototype system, that has the output of ten languages, was improved in view of the result of the first field trial. We added various expression forms found in the first trial data, and we also introduced the OOV detecting module using garbage models. The second field trial results showed that those improvements reduced misrecognitions due to the OOV inputs. The paper also describes the prototype of the bi-directional interpretation system between Japanese and Korean. We adapted the Hidden Markov Models (HMMs) made by the Japanese speech data to the Korean speech recognition system. The experimental results showed that the system works well for the fixed sentence recognition, even though its robustness in other tasks and different conditions remains to be discussed.

1. Introduction

Automatic speech interpretation has been a great challenge of the speech and language researchers, but they have not yet realized the system whose performance is good enough for consumers. The difficulties are due to the structure of the speech interpretation system. The typical system consists of the modules of speech recognition, machine translation, and speech synthesis, and the multiplication of the performance of each module makes the unsatisfying result. Many prototypes have been developed and worked well in the demonstration, but none of them became popular in the market. However, it seems that there is urgent necessity of an automatic interpretation system, even if it is not perfect. We thought that there is another way in developing the speech interpretation system. It would be acceptable to reduce the functions of the system, if the very high performance is guaranteed for the field data. As the minimum and realistic requirements, we accepted the following three limitations of the functions.

(1) Only the fixed sentences can be accepted.
(2) The interpretation is one-way.
(3) The target is focused on travelers.

The largest difficulty of the interpretation system originates in the recognition of the free input of the spontaneous speech. The first limitation raises the recognition accuracy easily. The second limitation is drawn naturally because the first utterance can be controlled by the initiating person of the dialogue, but the answer is difficult to control. The third limitation is to make a useful system within the above two limitations. There are many typical sentences for travelers, and even the one-way communication is helpful because they are guests.

The first prototype of our project was a portable interpreter[1], which was originally designed and manufactured. We applied the speech recognition middleware technologies for RISC microprocessors[2] to realize a

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

compact body of the instrument. However, we realized that most people would not like to buy and carry even the smallest machine. After all, we decided to use the cellular phones as the terminal of the speech interpretation system. Most people are always carrying their cellular phones, and then they do not need any initial cost to use our system.

The system is simple, but even the simplest system can be still difficult to use for the inexperienced people. To realize the truly user-friendly system, we have carried out the field trial of our system. Investigation of the users' attitude and the system performance would be necessary to improve the system. This paper focuses on the field trial tests and their results, and clarifies the feasibility and the points of improvement of our system.

The rest of the paper is organized as follows. Section 2 describes the overview of our previous work, that is about the development of the portable interpreter. Section 3 describes the detail of the developed system. In section 4 and 5, the condition and the results of the field trials are given. We also introduce the bi-directional interpretation system between Japanese and Korean in the section 6. Finally, conclusions and discussions are given in the last section.

2. Portable Interpreter

We had developed a prototype of a portable interpreter. Figure 1 shows the appearance of the portable interpreter. The prototype was equipped with a RISC microprocessor (60 MIPS), 8Mbyte SDRAM, and 16Mbyte flash memory. As the user interface, an LCD and six buttons were used. The size was 15cm (width), 6cm (height), and 3cm (thickness), and the weight was 180g including batteries.

This interpreter was based on keyword speech recognition and related sentence extraction. The dictionary was made of 30,000 words, and the error correction by Japanese syllable recognition assured the input of any of 30,000 words. Since it had an LCD, the speech recognition unit and the sentence extraction unit could display the N-best results. Even if the recognition rate for a large vocabulary was not so high, the N-best accuracy was enough to find the desired sentence.

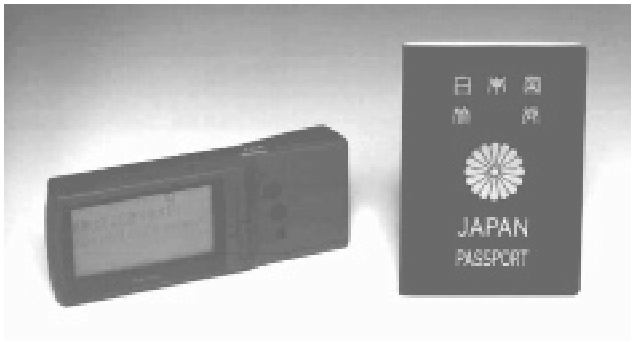


Fig.1 : Prototype of portable interpreter

3. Interpretation System Using Cellular Phones

The prototype of the portable interpreter was small and convenient, but there was still a problem that most people did not want to buy and carry a task-specified machine. The simplest answer to this problem is to use their cellular phones, although it produces some new problems.

Since the voice streaming technology for cellular phones has not yet become popular, the only way to realize the system based on the speech interface is to use the circuit switching service. When using the circuit switching service, the user can not use the LCD for confirmation or selection. Therefore, simple sequence from the input to the output is necessary, and then the recognition rate for single answer must be high enough to avoid the annoying error correction procedure.

To keep the recognition rate high, we have chosen the fixed sentence speech recognition, and we confines the size of the lexicon to 200 or smaller. We prepared 1,735 sentences in total, and divided them into 21 scenes, though some sentences appear in two or more scenes. The user interface is very simple. The user speaks one of registered sentences, and the server recognizes it, repeat the recognition result, and outputs the corresponding foreign sentence. The system architecture is shown in fig. 2.

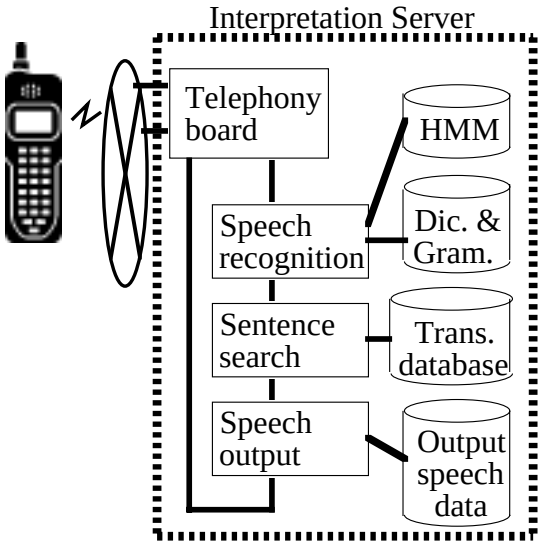


Fig.2 : System architecture

According to the specifications described above, we have developed the interpretation server system. The HMMs are made using the training data of the telephone voice, that matches the sampling condition of telephone. The environmental adaptation by the cepstrum mean normalization (CMN) is also applied to avoid the degradation of the recognition performance due to the various frequency characteristics. Moreover, the dictionary and the grammar are tuned to cope with the spontaneous input. The filler dictionary can handle some meaningless phrases before the sentence, and the grammar is written to accept some variations of the expression style that have almost same meaning. The dictionary also includes some command sentences to change the scene, to change the output language, to hear the guidance, or to quit the service.

Since the system is using fixed sentences, it is easy to expand the output language. We started with the Japanese-to-Korean system, and then added English, Chinese, German, French, Spanish, Portuguese, Italian, Russian, and Hindi as the output language. The expansion of the input language is rather difficult and remained as a future challenge, together with the problem of handling the wider variety of the answering utterances.

4. Results of the First Field Trial

It was easily predicted that our system would show high recognition rate for the laboratory data. Preliminary experiments have shown 95% to 98% accuracy for clean speech. However, our target is the users who are not experienced with the speech recognition system, and we could not estimate what would happen for such people. In order to collect the field data and evaluate our system, we have carried out field trial tests of our speech interpretation system.

Table 1 shows the detail of the first field trial test. Four lines were open to provide the service. The service itself was free of charge, but the user had to pay the calling charge. The phone number was announced by newspapers and other media. The detailed guidance was provided on the website, but the number of accesses to the website was much smaller than the number of accesses to the service itself. Therefore, it is reasonable to assume that many people used our system without detailed knowledge about the function of the system.

Figure 3 shows the classification of the input and corresponding action of the system. More details are shown in Table 2. The class 1, 2, and 3 correspond to the utterances that should be recognized correctly. The variation inputs, classified as 2, are only a little different from the registered sentences. There are some utterances not recorded completely, that are classified as 4. The class 5 includes out-of-vocabulary (OOV) utterances. Since it is theoretically impossible to recognize the utterances of (4) and (5), we defined "pure recognition rate" as (N(1) + N(2)) / (N(1) + N(2) + N(3)), where N(x) means the number of utterances in the class x. This value shows the performance of the recognition engine itself. We also defined "task completion rate" as (N(1) + N(2)) / (N(1) + N(2) + N(3) + N(4) + N(5)). This value implies the total performance of the user interface and the recognition engine.

The pure recognition rate for the all data is 83.9%. It is acceptable rate but lower than the preliminary experiment result. The error analysis showed that the various speaking style (lower power and faster speed) may

Table 1 : Details of field trial

Location	Tokyo (accessable from anywhere by telephone)
Language	from Japanese to Korean
Date	2000/11/14 - 2001/1/13
Time	24 hours
Lines	4
Charge	No service charge (calling charge only)
Contents	1682 sentences about travel (21 scenes)
Total access	6753

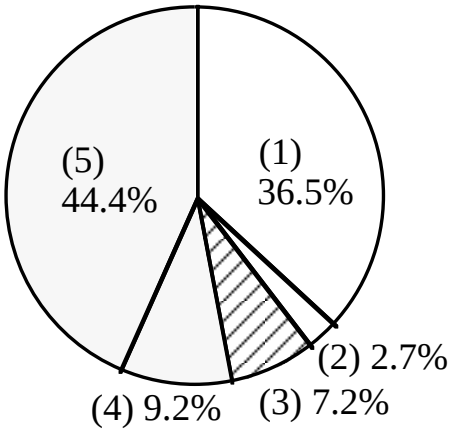


Fig. 3 : Classification of input and action
(1) Correct recognition for correct input
(2) Correct recognition for variation input
(3) Incorrect recognition for correct input
(4) Recording error of input speech
(5) Out of vocabulary

Table 2 : Details of recognition results

(%)

	recog. result	scene / input	Basic (default)	Other	Total
1	○	Correct	37.1	28.4	36.5
2	○	Variation	2.4	6.0	2.7
3	×	Correct	7.6	2.2	7.2
4	×	Rec. Error	9.3	8.1	9.2
5	×	OOV	43.6	55.1	44.4
Pure recog. rate			83.9	93.9	84.5
Task comp. rate			39.5	34.5	39.1

cause the performance reduction. However, the largest problem is the 43.6% OOV input. Since the announcement of the service by the newspaper could not describe the detail of the system, some people may have used the service without knowing that only the registered sentences can be translated. Because of this problem, the task completion rate is less than 40%. Even though it is impossible to recognize and translate these OOV sentences, at least the rejection of such sentences is strongly needed.

We analyzed the data by dividing them into the scene named "basic", and other scenes. Since the "basic" is the default scene, all the users first access to this dictionary. The user has to say "Change the scene" to move to another scene. Therefore, it is reasonable to expect that those who accessed to other scenes know well about the service. As shown in Table 2, the pure recognition rate for other scenes was 93.9%, which is almost same as the recognition rate for the laboratory data. The recognition rate for the experienced people is enough, and we concluded that the robustness for the inexperienced users is the point of the next improvement.

5. Results of the Second Field Trial

After the first field trial, we improved the system in view of the field data. We added some typical expression variations on the dictionary, to reduce the number of OOV inputs. We also introduced the OOV detecting module using garbage models. The OOV input is supposed to be matched with the sequence of five types of garbages, each of which is made of three random Japanese syllables.

Table 2 shows the details of the second field trial test. The second system has the output of ten languages such as English, Chinese, Korean, French, German, Spanish, Portuguese, Italian, Russian, and Hindi. The number of registered sentences had increased a little, but all the other condition is same.

Figure 4 shows the classification of the input and corresponding action of the system. Since the second system has the OOV reject module, two categories are added. One is the rejection for OOV inputs (class 7), and the other is the (wrong) rejection for correct inputs (class 3).

Table 3 : Details of field trial

Language	from Japanese to 10 lang. (ENG, FRE, CHN, etc)
Date	2001/04/26 - 2001/7/31
Contents	1735 sentences about travel (21 scenes)
Total access	4963

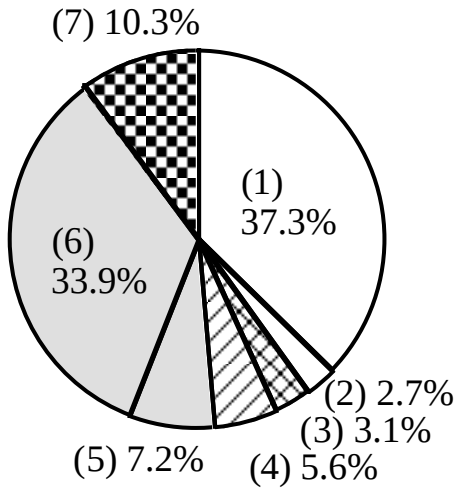


Fig. 4 : Classification of input and action
(1) Correct recognition for correct input
(2) Correct recognition for variation input
(3) Rejection for correct input
(4) Incorrect recognition for correct input
(5) Recording error of input speech
(6) Misrecognition of OOV input
(7) Rejection of OOV input

Table 4 : Details of recognition results

(%)

	recog. result	scene / input	Basic (default)	Other	Total
1	○	Correct	38.7	30.2	37.3
2		Variation	2.7	2.5	2.7
3	**Rej.**	Correct	3.2	2.5	3.1
4	×	Correct	5.8	4.8	5.6
5		Rec. Error	6.6	9.8	7.2
6		OOV	34.0	33.6	33.9
7	**Rej.**	OOV	9.0	16.6	10.3
Pure recog. rate			82.1	81.8	82.1
Pure recog. accuracy			87.7	87.2	87.6
Task comp. rate			41.4	32.7	39.9
Task comp. accuracy			47.2	40.0	45.7

The definition of the pure recognition rate was changed to (N(1) + N(2)) / (N(1) + N(2) + N(3) + N(4)), and the task completion rate was changed to (N(1) + N(2)) / (N(1) + N(2) + N(3) + N(4) + N(5) + N(6) + N(7)).

It is shown that the pure recognition rate was not improved from the first trial to the second. It is because there are some false rejections (class 3). It is noticeable that the recognition rates for the "other" scenes are lower than that for the "basic". It seems that there is a difference in users' attitude between the first and second trials, but the reason is not clear.

In order to show the efficiency of the OOV detecting module, we added the evaluation items by the recognition and task completion accuracies. The accuracy is calculated as the ratio of the correct results over the accepted utterances. The recognition accuracy and the task completion accuracy are higher than the recognition rate and the task completion rate in the first field trial, that shows the efficiency of the OOV detecting module. The current OOV detecting module might not be enough, therefore the improvement of this module would be quite important in the future system.

6. Japanese/Korean Bi-directional Interpretation System

The systems in the first and second field trial tests were one-way interpretation system from Japanese to other languages. However, it is necessary to understand the answer by the foreigner. We need many speech recognizers for many languages, but as the first step, we started with Japanese/Korean bi-directional system, because both languages seem to have similar pronunciation systems.

We have made a prototype of Japanese/Korean bi-directional interpretation system, by using only Japanese Hidden Markov Models. We expected that Japanese HMMs would basically work well, even though it can not distinguish some phrases including pronunciations that do not exist in Japanese. However, if the baseline performance is high enough, we can make Korean HMMs by the adaptation scheme. In such a case, the burden of the collection of the Korean training data would become smaller.

We have made the Korean-to-Japanese interpretation system, that has only 'basic' sentences of the field trial system. We made preliminary evaluation by only one speaker, and the recognition rate was about 90%. Moreover, the error analysis showed that the recognition errors were repeated in limited sentences, so if we can adapt some pronunciation in those sentences, the recognition rate would be improved easily.

7. Conclusions

We have developed the multilingual speech interpretation system using cellular phones. The system is based on the fixed sentence speech recognition scheme, therefore the recognition rate can be kept high. It should be noted that the addition of the output language is very easy, and we prepared the database of ten languages. In order to keep the robustness for the spontaneous input, the system is equipped with the grammar and the filler dictionary, that can handle the meaningless phrases and the various expression style of the same meaning. We have carried out the field trial test to evaluate the system performance and to collect the field data. The results showed that there were some recognition errors, especially for the unexperienced users, and the more important issue is the out-of-vocabulary (OOV) utterances. In the second field trial test, where the OOV detecting module was added, the recognition rate was not improved enough, but a lot of OOV inputs were rejected successfully. It means that the user interface could be improved by introducing the OOV detecting module. We also discussed about the bi-directional interpretation system, and developed the prototype of Japanese/Korean interpretation system. It showed the similarity of those two languages, and suggested that the same framework can be applied to the bilingual speech recognition system.

Acknowledgments

The authors would like to thank Dr. S. R. Kim and his staff in Samsung Advanced Institute of Technology, for their helpful discussions and offering Korean synthesized speech data used as the output of the first trial system.

References

[1] Obuchi, Y., Koizumi, A., Kitahara, Y., Matsuda, J., and Tsukada, T., “Portable Speech Interpreter Which Has Voice Input and Sophisticated Correction Functions,” *Proc. of Eurospeech '99*, 1999.

[2] Hataoka, N., Kokubo, H., Obuchi, Y., and Amano, A, “Development of Robust Speech Recognition Middleware on Microprocessor,” *Proc. of ICASSP ‘98*, pp.837-840, 1998.

Fusion of Output Scores on Language Identification System

Eddie Wong and Sridha Sridharan

Speech Research Laboratory, RCSAVT, School of EESE
Queensland University of Technology
GPO Box 2434, Brisbane QLD 4001, Australia
[ee.wong,s.sridharan]@qut.edu.au

Abstract

This paper investigates the fusion of multiple output scores generated using different parameterisation methods by the language identification system. A number of different fusion techniques including simple addition, simple multiplication, max rule and linear score weighting have been investigated. We have used a Gaussian mixture model based system to calculate the output scores. The universal background model adaptation technique has been incorporated to reduce the computation time on both training and testing phases. Five different parameterisation methods were generated by the system. They included mel-frequency cepstrum coefficients, linear prediction cepstrum coefficients, line spectrum pair frequencies, mel-cepstral coefficients and perceptual linear predictive coefficients. Results have shown that the fusion of mel-cepstral coefficients with perceptual linear predictive coefficients using the linear score weighting method improved the system accuracy from 75% to 79% on the 45 seconds test segment and from 66% to 71% on the 10 seconds test segment compared to the best independent scores without fusion. An error reduction of 5% and 7% respectively were achieved.

1. Introduction

Fusing data from different data sources has been shown to be capable of increasing a systems performance in many different tasks. Within speech technology, it has been successfully applied to gender identification [1], speaker identification and speaker verification system [2] and speech recognition systems [3]. Data fusion can be generally divided into input fusion and output fusion methods. Input fusion is simply the concatenation of different feature vectors for use by a classifier, while output fusion is the utilisation of the output from several classifiers to form the final scores. Language Identification (LID) typically process large amount of data in order to model a language's characteristics. Due to this, input fusion makes the input vector size becomes larger and thus reduce the efficiency of the system. Moreover, preliminary results show that output fusion performs better than the input counterpart. Therefore this paper focus the study on the output fusion method. The rest of the paper is organised as follows: Section 2 gives a review of the output score fusion techniques. Section 3 describes the LID test system and is followed by experimental results in Section 4. The conclusion is given in Section 5.

2. Data Fusion

The basic idea of data fusion is to combine different views or decisions generated by different experts (a classifier in this case) in an attempt to improve the discriminability of the overall system. Data fusion can be generally divided into input fusion and output fusion methods.

2.1. Input Fusion

Input fusion is simply the concatenation of different feature vectors into a single vector prior to processing by a classifier. A block diagram of an input fusion system is shown in Figure 1. The appending of energy, delta energy, delta and acceleration coefficients to the feature vector can be treated as a special case of input fusion.

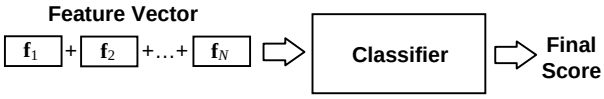


Figure 1: Input Fusion System. $\mathbf{f}_i$, $i = 1…N$, is the feature vector that generated by speech front-end *i*.

2.2. Output Fusion

Output fusion is the utilisation of the outputs from several classifiers to give a final score. Figure 2 shows a block diagram of such a system. Depending on the type of output the classifier generates, different styles of fusion can be performed. When continuous outputs such as likelihood scores are obtained from the classifier, both linear or non-linear combination methods can be applied. Although non-linear combinations like neural networks were shown to perform better than the linear methods [4,5], it has the disadvantage of a larger computational expense and the score modelling is more complex. Therefore only the linear combination and other simple related methods were investigated in this paper.

2.3. Output Fusion Strategies

As shown in Figure 2, the input feature vectors of each classifier was generated by different speech front-ends. This implies that different parameterisation methods were employed and therefore the output likelihood scores will have different dynamic range across the classifiers. Thus as a means of normalisation, we calculated the *a posteriori* probabilities instead of the likelihood scores as the output of each classifier. The *a posteriori* probability of model λ_i given a sequence of feature vectors $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_T\}$ is defined as

$$P(\lambda_i \mid \mathbf{X}) = \frac{P(\lambda_i)\, p(\mathbf{X} \mid \lambda_i)}{\sum_{l=1}^{L} P(\lambda_l)\, p(\mathbf{X} \mid \lambda_l)} \quad (1)$$

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

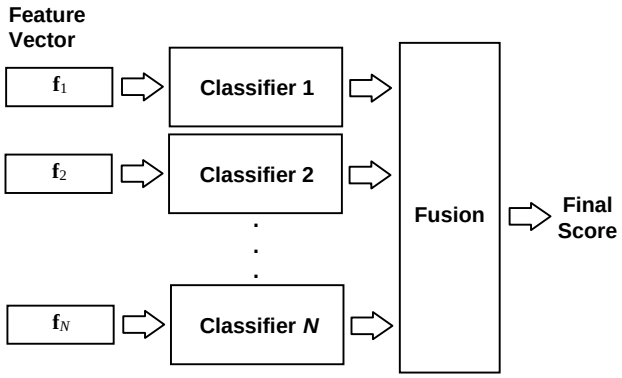


Figure 2: Output Fusion System. $\mathbf{f}_i$, $i = 1 \ldots N$, is the feature vector that generated by speech front-end *i*.

where L is the number of models for each classifier or the number of languages in this experiment and $P(\lambda_i)$ is the *a priori* probability of model λ_i. Also $p(\mathbf{X}|\lambda_i)$ is the joint likelihood of observations $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_T$ given the model λ_i. In our experiment we use the exponent of the expected frame-based natural-logarithm of the likelihood scores (see equation 8).

Four different linear output combination methods are investigated in this paper. They are defined as follows:

- Simple Addition (SA)

$$Z_{SA} = \sum_{i=1}^{N} Y_i \tag{2}$$

- Simple Multiplication (SM)

$$Z_{SM} = \prod_{i=1}^{N} Y_i \tag{3}$$

- Max Rule (MAX)

$$Z_{MAX} = \max_{i=1}^{N} Y_i \tag{4}$$

- Linear Score Weighting (LSW)

$$Z_{LSW} = \sum_{i=1}^{N} \alpha_i Y_i \tag{5}$$

where Z is the final output score, Y_i is the output score generated by classifier *i*, N is the number of classifiers and α_i is the score weighting for classifier output *i* such that

$$\sum_{i=1}^{N} \alpha_i = 1 . \tag{6}$$

3. Test System

3.1. Speech Parameterisation

The complete LID system includes five different parameterisation methods. They included:

- Mel-Frequency Cepstrum Coefficients (MFCCs) [6], MFCCs are one of the more popular filterbank based parameterisation methods used by researchers in the speech technology field. The advantage of applying the mel-scale is that it approximates the nonlinear frequency resolution of the human ear.
- Linear Prediction Cepstrum Coefficients (LPCCs) [7], LPCCs are Linear Prediction Coefficients (LPCs) represented in the cepstral domain. It have been widely used for a few decades and proven to be more robust and reliable than LPCs.
- Line Spectrum Pair frequencies (LSPs) [8], LSP frequencies were first introduced by Itakura as an alternative LPC spectral representation. They have been widely used in the speech coding domain.
- Mel-Cepstral coefficients (MCCs) [9], MCCs have been applied successfully to both speech coding and speech recognition. They do not have the disadvantage of LPCCs, which approximate speech linearly at all frequencies. Instead, the cepstrum is mapped to the mel-scale to model the auditory nonlinear frequency response.
- Perceptual Linear Predictive coefficients (PLPs) [10], PLPs have been widely used in speech recognition and have been shown to give good accuracy in different applications. Instead of modeling the spectrum of the speech as with LPCs, PLPs apply the LPCs' inverse filter to model the auditory spectrum.

3.2. Language Classifier

The testing system [11] uses Gaussian Mixture Models (GMMs) to model the characteristics of each target language. The GMM approach attempts to model the probability density function of a feature vector, $\mathbf{x}$, by the weighted combination of multi-variate Gaussian densities:

$$p(\mathbf{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}) \tag{7}$$

where λ is the model of a target language, M is the number of mixtures within the GMM, p_i is the weight applied to mixture *i* and $b_i(\mathbf{x})$ is the Gaussian density of mixture *i*.

During testing, a maximum likelihood classification is used to identify the language of the testing speech file. The average log-likelihood score of a model given a test speech file is calculated as:

$$\mathrm{E}[\log_e p(\mathbf{x} \mid \lambda)] = \frac{1}{T} \sum_{t=1}^{T} \log_e p(\mathbf{x}_t \mid \lambda) \tag{8}$$

where T is the number of feature vectors contained in the test speech file. The exponent of the expected score is used for an indicator of the *a posterior* probability.

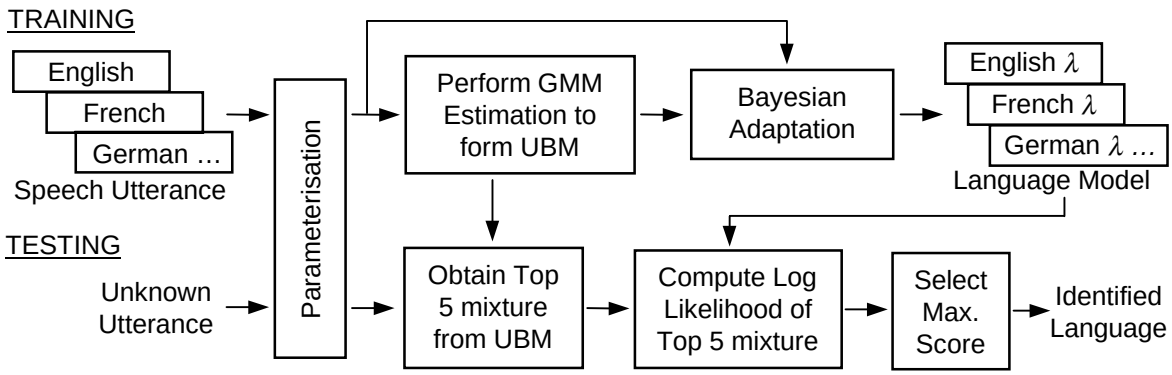


Figure 3: Block diagram of the LID system.

3.3. Universal Background Model

In order to reduce the training and testing time, the Universal Background Model (UBM) technique [12] is adapted in this system. An UBM in the LID case is a GMM representing the characteristics of all different languages. Instead of performing a full Expectation-Maximisation (EM) algorithm for training, the models of each langauge can be created by employing Bayesian adaptation [13] from the UBM. Therefore significant amounts of time are saved in the training of each language model.

Reynolds [12] has found that only a few of the mixtures of a GMM contribute significantly to the likelihood score. The adapted model of each language will share a certain correspondence with the UBM, since each model is adapted from it. Therefore the average log-likelihood score of each language model can be calculated by only taking into account the most significant mixtures. According to the second property mentioned above, these significant mixtures can be obtained by selecting the mixtures from the UBM that have the highest score. In this way, the computation required for testing can also be reduced significantly.

A block diagram of the system is shown in Figure 3. The main advantage of this sytem is that the complexity is low compared to other LID systems such as the Phoneme Recognition followed by Language Modelling performed in Parallel (PRLM-R) system [14] and the large vocabulary continuous speech recognition based system [15]. Thus this system is suitable to operate in real time even with the inclusion of the output score fusion technique. Another advantage provided by this system is that no transcriptions of training data are required. This makes the implementation and adaptation to new languages relatively easy.

4. Experiments and Results

The experiment was performed using the Oregon Graduate Institute Telephone Speech (OGI-TS) Corpus [16] which included the following 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The 1994 National Institute of Standards and Technology (NIST) LID evaluation specification was used as a guideline for extracting the training and testing data to perform the experiment. Both the training and extended data from the corpus were used for creating the UBM and adapting language models. The 45 second (187 test segments) and 10 second (625 test segments) test data available for testing.

Each feature vector was extracted at 10ms intervals using a 32ms window, with each frame of speech being pre-emphasised by the first order difference equation $s'_n = s_n - 0.97 s_{n-1}$ except for PLP coefficients. A Hamming window was then applied to the speech frame. The GMMs used 512 mixtures to model each language. The specifications for each parameterisation are: MFCC (26 filters, 12 cepstral coefficients), LPCC (14th order LPC, 12 cepstral coefficients), LSP (12 coefficients), MCC (14th order LPC, 256th order LPCC, 12 cepstral coefficients) and PLP (17 filters, 5 cepstral coefficients). Delta energy, delta and acceleration coefficients were appended to the feature vectors with mean and variance normalisation applied. The accuracy of these baseline systems is shown in Table 1. The MCC and PLP performed slightly better than other methods. The LSP features tend not to perform as well as the cepstral based features..

Table 1: LID Results comparing baseline systems. Results are in percentage accuracy.

Method	45s	10s
MFCC	66.8	62.4
LPCC	69.5	63.0
LSP	62.0	58.1
MCC	74.3	65.8
PLP	74.9	66.2

4.1. Compare Data Fusion methods

In this evaluation, we focused on the linear fusion methods. The methods studied are SA, SM, MAX and LSW. The results of the comparison are shown in Table 2. The number of classifiers, N, used for this experiment is limited to two and the *a priori* probability, $P(\lambda)$, of model λ, is set to equal probabilities $P(\lambda) = (1 / 11)$.

All the results shown in Table 2 are consistent with the accuracy of Table 1. The best result is the fusion of PLP (α = 0.8) with MCC (α = 0.2) using LSW with an accuracy of 79% for the 45 second test and 71% for the 10 second test. The LSW method has the highest improvement amongst the tested fusion methods. The performance of SA and SM are approximately the same. Both methods have a slight improvement in accuracy after fusion. The MAX method does

Table 2: LID Results comparing linear fusion methods. Results are in percentage correct. α is the score weighting applied to the first parameterisation method under Fusion column.

Fusion	SA		SM		MAX		LSW			
	45s	10s	45s	10s	45s	10s	45s	α	10s	α
PLP-MCC	75.9	70.6	75.9	70.6	74.3	66.2	**79.1**	0.80	**71.4**	0.77
PLP-LPCC	74.9	69.0	74.9	68.6	69.5	63.4	78.6	0.75	70.1	0.71
PLP-MFCC	70.1	65.0	70.1	64.8	67.9	62.2	76.5	0.93	67.7	0.86
PLP-LSP	69.0	63.2	69.5	63.5	63.1	58.9	75.4	0.92	67.5	0.91
MCC-LPCC	72.2	67.4	72.2	67.2	74.3	67.4	74.9	0.84	67.8	0.69
MCC-MFCC	72.7	65.8	72.7	65.6	73.8	64.8	75.4	0.89	67.2	0.67
MCC-LSP	72.7	65.4	73.3	65.4	70.1	64.5	75.4	0.8	67.5	0.77
LPCC-MFCC	70.1	64.0	70.1	64.3	67.9	63.8	71.1	0.68	65.1	0.71
LPCC-LSP	69.0	63.0	69.0	63.0	65.2	61.4	70.1	0.78	64.0	0.69
MFCC-LSP	66.8	62.1	67.4	61.8	62.6	61.0	67.9	0.58	63.2	0.83

not show any improvement in accuracy. This may indicate that normalising the output score by the *a posteriori* probability may not be appropriate. This may also attributed to one type of parameterisation expert dominating the others.

The results shown in the LSW column of Table 2 are the best accuracies obtained by exhaustively searching through all the score weighting combinations. The search is done by stepping through the α value (of equation 5) from 0.0 to 1.0 with each step incremented by 0.01. This searching approach is only possible when the correct results are available. Unfortunately, these results do not indicate the robustness of the fusion approach. However it is employed in this experiment to show that a proper choice of score weighting is important to the LSW method with different speech parameterisation front-ends. In other applications, the optimised score weighting can be obtained by performing the aforementioned searching approach on the output scores generated from a set of development test data.

4.2. Varied Number of Classifiers

In the last experiment we found that LSW yielded the best improvement for all different parameterisation combinations when the number of classifier was set to two. In this experiment the fusion method is fixed at LSW and the number of classifiers, *N*, are varying. Table 3 shows the results of this experiment. Again the *a priori* probability, $P(\lambda)$, of model λ, is set to equal probable and the optimised score weighting searching approach was applied to the LSW results.

The result shows that as the number of classifiers increase, the accuracy remains the same or increase. This is consistent with information theory. However, from the PLP-MCC row and PLP-MCC-MFCC row, the amount of improvement is not significant. But comparing the PLP-MFCC row with the PLP-MCC-MFCC row, a more significant improvement was obtained. It should be noted that the information provided by different parameterisation methods is not entirely independent. From Table 1, PLP and MCC have the highest accuracy, which means that they can provide more discriminative information than other methods. Fusion of PLP with MCC thus covered most of the information that other feature types were capable of providing. Further increasing the number of classifiers in the fusion stage using similar features will only add a small amount of information to the final decision. Therefore by choosing the correct parameterisation methods for the output score fusion LID system, the number of classifiers required to achieve the optimal performance can be reduced.

Table 3: LID Results comparing varies number of classifier. Results are in percentage correct.

Fusion	LSW	
	45s	10s
***N* = 2**		
PLP-MCC	79.1	71.4
PLP-LPCC	78.6	70.1
PLP-MFCC	76.5	67.7
PLP-LSP	75.4	67.5
MCC-LPCC	74.9	67.8
MCC-MFCC	75.4	67.2
MCC-LSP	75.4	67.5
LPCC-MFCC	71.1	65.1
LPCC-LSP	70.1	64.0
MFCC-LSP	67.9	63.2
***N* = 3**		
PLP-MCC-LPCC	79.7	71.8
PLP-MCC-MFCC	79.1	71.5
PLP-MCC-LSP	79.1	71.4
PLP-LPCC-MFCC	78.6	70.4
PLP-LPCC-LSP	78.6	70.1
PLP-MFCC-LSP	76.5	68.0
MCC-LPCC-MFCC	75.4	68.3
MCC-LPCC-LSP	75.9	68.5
MCC-MFCC-LSP	75.4	68.5
LPCC-MFCC-LSP	71.7	65.8
***N* = 4**		
PLP-MCC-LPCC-MFCC	79.7	71.8
PLP-MCC-LPCC-LSP	79.7	71.8
MCC-LPCC-MFCC-LSP	75.9	69.0
***N* = 5**		
PLP-MCC-LPCC-MFCC-LSP	79.7	71.8

5. Conclusions

This paper investigated several techniques of output score fusion using different parameterisation methods for a

language identification system. Experiments are conducted on the 1994 NIST Langauge Identification Evaluation which is based on the OGI-TS Corpus. The fusion of mel-cepstral coefficients with perceptual linear predictive coefficients using the linear score weighting method improved the system accuracy from 75% to 79% on the 45 seconds test segment and from 66% to 71% on the 10 seconds test segment compared to the scores without fusion. An error reduction of 5% and 7% respectively were achieved. These result show that with the help of output fusion, the GMM-UBM LID system can perform comparably well against other more complex systems. It also has the advantage of a real time operating capability and no requirement of transcribed training data.

6. Acknowledgements

This work is sponsored by a research contract from the Australian Defence Science and Technology Organisation (DSTO). The author would like to thank Jason Pelecanos for valuable advice provided.

7. References

[1] Slomka, S. and Sridharan, S., "Automatic Gender Identification Optimised for Language Independence," *IEEE TENCON*, pp. 145-148, 1997.

[2] Genoud, D., Gravier, G., Bimbot, F., and Chollet, G., "Combining methods to improve the phone based speaker verification decision," *Internaional Conference on Spoken Language Processing*, vol. 3, pp. 1756-1760, 1996.

[3] Tibrewala, S. and Hermansky, H., "Sub-Band Based Recognition of Noisy Speech," *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. 1255-1258, 1997.

[4] Kittler, J., "Combining classifiers: A theoretical framework," *Pattern Analysis and Applications*, vol. 1, pp. 18-27, 1998.

[5] Sharma, S., Vermeulen, P., and Hermansky, H., "Combining information from multiple classifiers for speaker verification," *Speaker Recogniton and its Commercial and Forestic Applications*, 1998.

[6] Davis, S. B. and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, pp. 357-366, 1980.

[7] Markel, J. D. and Gray, A. H., *Linear prediction of speech*. New York: Springer-Verlag, 1976.

[8] Itakura, F., "Line Spectrum Representation of Line Predictive Coefficients of Speech Signals," *Journal of the Acoustical Society of America*, vol. 57, pp. S35, 1975.

[9] Tokuda, K., "Speech Signal Processing Toolkit," http://kt-lab.ics.nitech.ac.jp/~tokuda/SPTK/, 2000.

[10] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, pp. 1738-1752, 1990.

[11] Wong, E., Pelecanos, J., Myers, S., and Sridharan, S., "Language identification using efficient Gaussian Mixture Model analysis," *Australian International Conference on Speech Science & Technology*, pp. 78-83, 2000.

[12] Reynolds, D. A., "Comparison of background normalization methods for text-independent speaker verification," *Eurospeech*, vol. 2, pp. 963-966, 1997.

[13] Gauvain, J. L. and Lee, C. H., "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291-298, 1994.

[14] Zissman, M. A., "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 31-44, 1996.

[15] Mendoza, S., Gillick, L., Ito, Y., Lowe, S., and Newman, M., "Automatic language identification using large vocabulary continuous speech recognition," *International Conference on Acoustics, Speech and Signal Processing*, Atlanta, GA, USA, vol. 2, pp. 785-788, 1996.

[16] Muthusamy, Y. K., Cole, R. A., and Oshika, B. T., "The OGI multi-language telephone speech corpus," *International Conference on Spoken Language Processing*, vol. 2, pp. 895-898, 1992.

This page has been deliberately left blank

Page intentionnellement blanche

Multilingual Processing for Operational Users

Keith J. Miller, Florence Reeder, Lynette Hirschman, David D. Palmer

The MITRE Corporation
7515 Colshire Drive
McLean, VA 22102-7508
{keith, freeder, lynette, palmer}@mitre.org

Abstract

This paper describes multilingual technology projects currently being undertaken in conjunction with the NATO BICES (Battlefield Information Collection and Exploitation) organization. First, we describe the basis of the multilingual processing for these projects, the CyberTrans machine translation environment, an operational system that enables the use of machine translation (MT) by intelligence analysts [1]. We will briefly describe the impetus behind the development of CyberTrans as well as the system design and implementation. Next, we will discuss the operational pilot installation of CyberTrans on the BICES network. Finally, we will present some potential multilingual technology pilot experiments for BICES. These future technologies will enable NATO to meet the challenges of its inherently multilingual user community and will pave the way for interoperability across language barriers in the future.

1. Introduction

The challenge of automatic processing of language data from multiple languages is increasingly diverse and problematic. In addition to the wealth of work that has been done on the English language, "foreign language" processing needs are increasing, in part because of the changing conditions and needs in the world. Traditionally, users could focus on just a few foreign languages and a limited number of sources of foreign language materials. As we begin the 21st century, users of online materials are faced with having to process, utilize and exploit documents that may be in one of many languages or a combination of languages. It is not feasible to expect a given user to know all of the languages related to their topic of research. It is equally unrealistic to expect to have on-demand translators available in every language whenever they are needed. Because of the expanding need, tools are being developed to automate the use of foreign language materials.

A key component of many multilingual applications is machine translation (MT). A common vision for machine translation is as a small part of a larger process that is partly or completely automated. For many users, this does not mean having to work with yet another tool and yet another interface, but a nearly invisible companion that incorporates translation and necessary support technologies. One such system, the United States Army Research Lab (ARL) FALCon system [1,2], combines scanning, optical character recognition (OCR), translation and filtering into a single process. Another view of this is the DARPA Translingual Information Detection, Extraction and Summarization effort (TIDES) [3]. TIDES represents the pinnacle of information access and is a significant challenge for MT. MT supports the translingual aspects of the effort and here can be viewed as an embedded tool that facilitates other technologies. Finally, the integration of MT into the process for intelligence analysis serves as the basis for the CyberTrans project [4].

2. CyberTrans

The first incarnation of CyberTrans grew out of a demonstration that machine translation could be useful in the intelligence analysis process. As a result of a survey of MT technology (Benoit et al, 1991), it was believed that MT was ready for incorporation into an operational environment. Questions remained, however, about which commercial off the shelf (COTS) or US government off the shelf (GOTS) MT engines to use, and how to make them accessible in a user-friendly manner. Thus, CyberTrans itself is not machine translation per se, but it is a way to make machine translation (both COTS and GOTS) tools available to a wide range of linguists and analysts. It incorporates the Systran family of MT systems, which provides several language pairs (German, French, Spanish, Portuguese, Italian, Russian, Serbo-Croatian, Ukrainian, Chinese, Japanese and Korean to English) free to for US government use. In addition, CyberTrans incorporates the Globallink (Lernout and Hauspie) tools, which provide German, French, Spanish, Russian to English translation, as well as the GOTS product Gister. Initially, CyberTrans was designed as a wrapper around MT systems in Unix environments. Based on a client-server architecture, it provides a common user interface to its multiple translation engines. The server software interacts with the translation engines, controlling the flow of the translations, and the client software handles the user end of the transaction. Historically, four clients were provided: e-mail, web, FrameMaker, and command line. By providing translation through these media, users could translate documents in a familiar interface without having to be concerned with differences between translation products.

Shortly after the fielding of the initial prototype, the need for additional language services to accompany translation became apparent. The "real world" data sent to the translation engines pointed out the differences between translation in an interactive environment and translation in an embedded, automated environment. Interactive translation is much more forgiving of low quality input data while automated processing must handle issues arising from data quality on the fly. Given the assumption that the user has no control over the production of the source document, and hence no control over the quality of that document, a series of pre- and post-processing tools were incorporated into CyberTrans, thus transforming it from a user-interface wrapper to a value-added machine translation environment.

Paper presented at the RTO IST Workshop on "Multilingual Speech and Language Processing", held in Aalborg, Denmark, 8 September 2001, and published in RTO-MP-066.

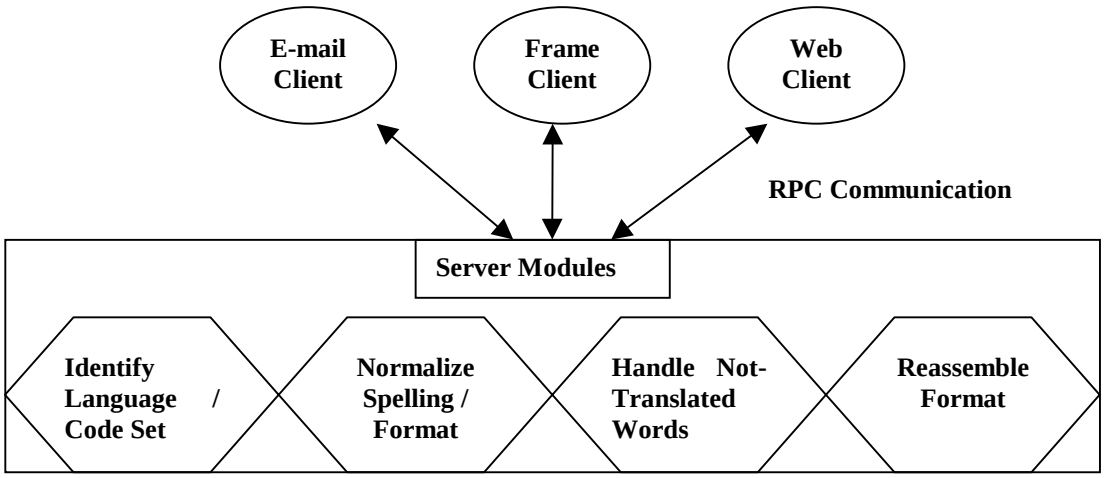


Figure 1: Original CyberTrans Architecture

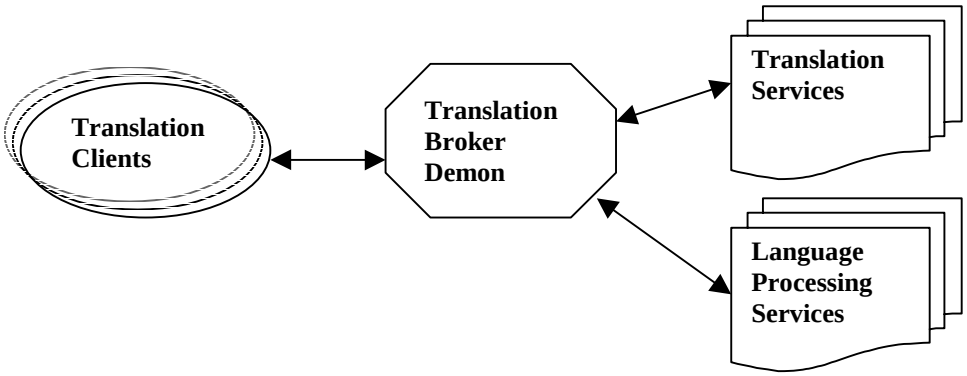


Figure 2: Updated CyberTrans Architecture

Initially, these pre- and post-processors were included in the functional architecture as depicted in Figure 1.

The server portion of the shell incorporates a) language and code set identification; b) language and code set conversion; c) limited spell checking (particularly diacritic reinsertion); d) format preservation. The data flow follows the following steps. Upon being submitted to CyberTrans, a document proceeds through steps a-d and is then passed on to the appropriate translation engine. The results are re-packaged, and the results are sent to the client for presentation to the user. As mentioned above, a number of clients are available: an e-mail client (send an e-mail to a specific address, get a translation back); a web-client (cut and paste or provide a URL); a command-line client and an API. The API allows integration of CyberTrans into a number of processes including word processing packages (FrameMaker; LotusNotes) and other applications, such as the TrIM, MITAP, and Open Sesame applications, discussed in Section 5.

The increased complexity caused by the addition of these language tools caused a necessary re-design of the architecture from a client-server model to an enterprise service model. This model is characterized by an open architecture of loosely coupled modules performing services for multiple applications. In this architecture, daemon processes broker translations. A request for translation is passed to the system by a client program, and a translation plan consisting of a series of translation-related services is created. Each service is requested from the responsible system object, and the resulting translation is passed back to the client programs. Implemented in a combination of C++ and Java, the new version represents a service-oriented architecture. Figure 2 shows this updated view of the architecture.

Language processing services include language/code set identification; code set conversion; data normalization, including diacritic reinsertion and generalized spell checking; format preservation for Hyper-Text Mark-up Language (HTML) documents; not-translated word preservation and logging, and others. The available clients remain e-mail, Web and FrameMaker. Platforms include both Unix and PC for clients, with the capability to incorporate PC-based translation tools as part of the service.

Many of the modifications and improvements in the system came as a direct result of the deployed of CyberTrans in an operational environment with a steadily-increasing user base. As can be seen in figure 3, between April 1998 and May 2000, CyberTrans experienced over 600% growth in monthly usage. In Section 3, we turn to lessons learned as a result of having

an operational MT capability, running 24 hours a day, 7 days a

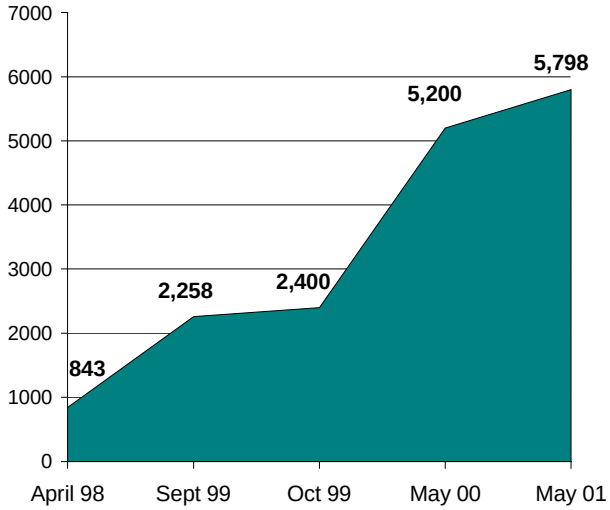


Figure 3 CyberTrans Usage Statistics

week, servicing over 4500 translation requests per month

3. Operational Machine Translation on BICES with CyberTrans

Since its initial installation, CyberTrans has been running around the clock at a US government site. It is currently available on a secure internal network and processes between 4500 and 6000 MT requests per month in up to 30 different languages. While not approaching the "Star Trek" vision of perfect MT quality, it does provide useful translations for the purpose of scanning and filtering by intelligence analysts. Since its initial deployment as an operational system, we have learned much in the realm of real world translation. One obvious lesson is that the better the quality of machine translation is, the more invisible translation is to the user. Since our users range from those who cannot identify the language of the document they wish to process to linguists trying to speed up their work, we are familiar with a wide range of issues pertaining to users with respect to MT language processing.

This success led to interest in CyberTrans from other organizations in which the US government is involved. In addition to the initial system deployment, we have recently installed a version of CyberTrans on the NATO BICES (Battlefield Information Collection and Exploitation) network. The BICES network facilitates the coordination of battlefield intelligence gathering among the NATO nations. BICES is an ideal candidate for a pilot CyberTrans for two reasons. First, although French and English remain the official operational languages, NATO has 19 member countries representing upwards of 15 languages and has much to gain from an operational installation of usable MT. Second, access to native speakers of so many languages is a rich source of feedback for continued improvement of the CyberTrans environment and its component technologies.

Following from these characteristics, the following goals were set forth for the pilot installation of CyberTrans on BICES:

- Provide a standing MT capability on BICES that can be used by BICES users via the BICES Backbone Network (BBN) and can be improved over time.
- Provide feedback from BICES users to MITRE's DARPA TIDES team throughout this process.

This installation was realized with the support of the DARPA TIDES program.

Because of the quality of the output produced by state-of-the-art MT systems, there must be a balance between "selling the technology" and managing (potential) users' expectations in operational environments. In short, MT is a useful technology if it is not oversold, but also not undersold. For the BICES installation in particular, the following questions were focused on:

- Who is the user or customer?
- What are the user's requirements?
- How does MT fit into the overall user's process?
- What is the price/risk of miscommunication?
- What are the user's expectations?

For BICES, the initial users are volunteer native speakers of the various languages supported by the MT pilot. These users provide valuable feedback as to the utility of CyberTrans and on opportunities for improvement. After this initial phase, CyberTrans is made available to a wider community of users, who primarily use CyberTrans for purposes of information assimilation, such as understanding documents published in a language other than their native language. It is hypothesized that a small number of users will also use CyberTrans to assist them in producing certain documents for which there is an English-language reporting requirement. In both cases, users access machine translation on an as-needed basis, from their desktops, to translate documents that they already have in electronic form. Given the users' understanding that this is a pilot application, the output of the translation is not expected to be perfect. Furthermore, with the same limitation in mind, it is expected that users be highly skeptical regarding any seemingly suspicious translation output, which will hopefully mitigate the risk of any miscommunication. Finally, the BICES support team and initial users are briefed on the prospects and limitations of MT technology in general, in an effort to manage users' expectations of the technology. This information is to be passed on to the end users of the CyberTrans application. Additionally, messages about the realistic use of machine translation are prominently displayed in CyberTrans' web page interface.

With these answers to the guiding questions in mind, CyberTrans was installed on the BICES application test facility in May 2001 and was given a limited release on the operational network (the BBN) in late June. In July, the application was fully released and could be accessed by all BICES users. This initial implementation includes translation from Russian, German, French, Spanish, Italian, and Portuguese into English. Users have enthusiastically been providing feedback bearing on the translation quality, the pre- and post-processing facilities, and on the user interface, all of which will be used to enhance future versions of the system. In addition, users of the Portuguese to English translation

facility have been particularly positive in their response and have requested the addition of translation in the reverse direction.

4. The Value of User Feedback: Tailoring CyberTrans

The ultimate success of automated language processing applications depends on the breadth, depth, and overall quality of the lexical information being used in the systems. Accordingly, the highest portion of the cost of providing an MT capability reflects the amount of lexicography – or domain specialization – that must be done. It can total up to 70% of the cost of an MT engine and represents the greatest source of user dissatisfaction. In addition, many applications require specialized lexical repositories that reflect unique domains such as military, legal, scientific and medical terminology. We must find ways to update lexicons intelligently, using sources such as dictionaries, working aids, specialized word lists and other information reservoirs to provide broad vocabulary coverage. Our principal current approach is to record the list of words that do not translate and automate the handling of these.

Now that CyberTrans is in the operational phase of its pilot installation, a process will be put in place whereby logs of not-translated words are transferred back to the development team for use in updating the MT lexicons. Since CyberTrans can incorporate various MT engines, it is an interesting problem that different translation engines encode lexical entries in different ways, such that sharing lexicon entries between translation capabilities is problematic. We are working on lexicon service bureau (LSB) research designed to facilitate the sharing of lexical materials. One part of this is the automatic extraction of lexical entries from on-line, machine-readable dictionaries. Another part is the analysis of not-translated words. Each advance in this realm increases the overall quality of the output produced by machine translation systems.

In addition to domain-specific jargon and acronyms, proper names ("named entities") represent a complication for translation. Users of the pilot installation of CyberTrans on BICES have remarked that this is a particular problem for the Spanish to English translation output. In a recent test conducted on 100 news articles translated from Spanish to English, 2600 named entities were found. Two translations produced by human translators agreed on the "proper" translation of names only about 90% of the time. This, along with the user feedback, is a further indication that this phenomenon needs to be studied more in order to determine how to best handle these proper names in MT.

5. What the Future May Hold

5.1. TIDES Tools

In addition to being used to produce translated documents as an end product, CyberTrans is also being utilized in the DARPA Translingual Information Detection, Extraction, and Summarization (TIDES) program. Its role in TIDES is to support the goal of information access from a variety of sources in multiple languages. DARPA's TIDES program is working with NATO partners to integrate language processing and translation technology into future intelligence networks.

This effort is supported in part by an Integrated Feasibility Experiment led by The MITRE Corporation, with a focus on automatically filtering, extracting and summarizing information about the outbreak and spread of natural and man-made disease. This technology integration and application effort called the MITAP (MITRE Text and Audio Processing) System uses the CyberTrans embedded machine translation capability to translate information from languages such as Portuguese, Chinese, Russian, or Spanish into English. MT is a key component within TIDES research and has been a top requirement of the NATO member nations' priority requirements. In addition to the MT technology itself, the BICES technology team has expressed some interest in the other component technologies contained in the TIDES effort. It has been suggested that due to its multinational nature, the BICES network would be a rich operational testbed for some of the technologies that make up the "IDES" portion of TIDES.

5.2. TrIM

In coalition operations such as those supported by NATO, participants with a wide range of native languages must be able to coordinate their efforts. Collaboration between coalition partners currently relies on the ability to settle on a single common language for all participants. This arrangement creates communication bottlenecks, and is likely to work less well at lower echelons and in the field than it does at higher echelons or in command centers. Additionally, the information that must be communicated (including source documents in native languages) can be in multiple languages and specialized domains.

From a linguistic point of view, translingual collaborative environments represent a new and exciting area of research. From an operational standpoint, emerging technologies in translingual collaboration enable us to envision environments in which nations are able to collaborate across language barriers in ways never before thought possible. Translingual collaboration produces challenges that have heretofore not been addressed. Linguistic analysis for purposes of natural language processing applications has typically fallen into one of two camps: spoken interaction between two or more active participants and written interaction for information dissemination or assimilation. Each of these has unique characteristics, and there is often little overlap between the two. Collaborative environments, however, yield a new form of interaction, one which has some characteristics of spoken interaction and some characteristics of written interaction. This mix presents unique challenges for MT-mediated collaborative computing.

Plans are currently in the works for implementation of a Translingual Instant Messenger (TrIM) prototype on the BICES network. As part of this pilot effort, we will be able to collect data highlighting the multilingual challenges faced by collaborative environments, including the unique interaction style, the specialized needs of the interactions, the difficulty of analyzing actual language use, and the inherent difficulties of MT use in an interactive environment. These studies will enable the improvement and tailoring of MT technology for this specialized use; it is anticipated that such improvements will take place along several dimensions and will include data normalization, lexical improvements, and syntactic enhancements both in the analysis and generation phases of translation.

We have demonstrated the appeal of a Translingual Collaborative tool with Translingual Instant Messenger (TrIM) but this gives us only a view to the future, not the pieces necessary to make that future a useful reality. To make the translingual sharing of a reality, we need to develop tools for capturing, analyzing and enabling translingual information sharing. For instance, in TIDES we need a way to rapidly acquire domain-specific terminology and make it usable to the automated processing tools. In TrIM, the translation capability must be more reflective of the language style and terminology that is actually required for collaborative, coalition work.

5.3. Open SESAME

Open SESAME is the BICES intelligence production and discovery system. Key to sharing information between the nations is a “library card” representing product information in the form of metadata. This system is the bedrock for information exchange between the seventeen BICES nations. The BICES nations produce all their intelligence in their native language, which is far too much to reasonably translate manually. The library card concept allows nations to publish products in their native language, capturing the document metadata in English. Discovered documents deemed important may be passed on for translation.

The planned incorporation of CyberTrans with Open SESAME will carry this concept and process to a higher level. During the submission process, it will be possible to pass metadata fields, such as title and summary, through CyberTrans, automatically generating an English equivalent for the library card. Similarly, intelligence researchers may then submit native language title and summary searches, which will then be passed through CyberTrans, to query the Index Server’s English metadata tags for relevant intelligence products. Finally, free texts documents, may subsequently be passed to CyberTrans for full machine translation.

6. Conclusion

Due to its multinational character, the NATO BICES agency is a natural proving ground for multilingual technologies. It is also one of the organizations that stands to gain the most from the successful implementation of such technologies and from improvements that result from live pilot installations such as the one described in this paper. Working in conjunction with the BICES agency, we can make valuable multilingual technologies available to those who need them the most, while at the same time gathering crucial data that will enable researchers and developers to push those technologies to the next level of performance.

7. References

[1] Voss, C. R., Van Ess-Dykema, C., “When is an Embedded MT System "Good Enough" for Filtering?”, *Proceedings of Embedded Machine Translation Systems - Workshop II, ANLP-NAACL2000*, Seattle, May 2000.

[2] http://rpstl.arl.mil/ISB/falcon.htm

[3] Hirschman, L., Concepcion, K., Damianos, L., Day, D., Delmore, J., Ferro, L., Griffith, J., Henderson, J., Kurtz, J., Mani, I., Mardis, S., McEntee, T., Miller, K., Nunan, B., Ponte, J., Reeder, F., Wellner, B., Wilson, G., Yeh, A., “Integrated Feasibility Experiment for Bio-Security: IFE-Bio A TIDES Demonstration”, *Proceedings of HLT 2001 Human Language Technology Conference*, James Allan, ed., San Diego, March 2001.

[4] Reeder, F., “At Your Service: Embedded MT as a Service”, *Proceedings of Embedded Machine Translation Systems - Workshop II, ANLP-NAACL2000*, Seattle, May 2000.

This page has been deliberately left blank

Page intentionnellement blanche

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's References	3. Further Reference	4. Security Classification of Document
	RTO-MP-066 AC/323(IST-025)TP/21	ISBN 92-837-1102-5	UNCLASSIFIED/ UNLIMITED

5. Originator Research and Technology Organisation
North Atlantic Treaty Organisation
BP 25, F-92201 Neuilly-sur-Seine Cedex, France

6. Title
Multilingual Speech and Language Processing

7. Presented at/sponsored by
the Information Systems Technology Panel (IST) Workshop held in Aalborg, Denmark, 8 September 2001.

8. Author(s)/Editor(s)	9. Date
Multiple	April 2003
10. Author's/Editor's Address	**11. Pages**
Multiple	84

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

Automated language processing
Automated Speech Recognition (ASR)
Digital techniques
Human factors engineering
Information systems
Intelligibility
Languages
Linguistics
Machine translation
Multilingualism
Pattern recognition
Precision
Speech analysis
Speech processing
Speech recognition
Speech technology
Terminology
Translating
Voice communication
Voice recognition
Waveforms
Word recognition
Words (language)

14. Abstract

This volume contains the 12 papers, presented in 4 sessions at the Information Systems Technology Panel Workshop held in Aalborg, Denmark on 8th September 2001.

The papers presented covered the following headings:

- N4 Corpus and Speaker Indentification
- Non-Native Speech
- Speech Recognition
- Language Identification and Multilingual Applications

This page has been deliberately left blank

Page intentionnellement blanche

NORTH ATLANTIC TREATY ORGANISATION

RESEARCH AND TECHNOLOGY ORGANISATION
BP 25
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
Etat-Major de la Défense
Département d'Etat-Major Stratégie
ACOS-STRAT-STE – Coord. RTO
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
DSIGRD2
Bibliothécaire des ressources du savoir
R et D pour la défense Canada
Ministère de la Défense nationale
305, rue Rideau, 9e étage
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)
Defence Industry & Research General Directorate
Research Directorate
Fakinos Base Camp
S.T.G. 1020
Holargos, 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 di Documentazione Tecnico-Scientifica della Difesa
Via XX Settembre 123a
00187 Roma

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

PAYS-BAS
Royal Netherlands Military Academy Library
P.O. Box 90.002
4800 PA Breda

POLOGNE
Armament Policy Department
218 Niepodleglosci Av.
00-911 Warsaw

PORTUGAL
Estado Maior da Força Aérea
SDFA - Centro de Documentação
Alfragide
P-2720 Amadora

REPUBLIQUE TCHEQUE
DIC Czech Republic-NATO RTO
VTÚL a PVO Praha
Mladoboleslavská ul.
Praha 9, 197 06, Česká republika

ROYAUME-UNI
Dstl Knowledge Services
Kentigern House, Room 2246
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
Acquisitions
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)

NORTH ATLANTIC TREATY ORGANISATION

RESEARCH AND TECHNOLOGY ORGANISATION
BP 25
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 Organisation (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
Etat-Major de la Défense
Département d'Etat-Major Stratégie
ACOS-STRAT-STE – Coord. RTO
Quartier Reine Elisabeth
Rue d'Evère, B-1140 Bruxelles

CANADA
DRDKIM2
Knowledge Resources Librarian
Defence R&D Canada
Department of National Defence
305 Rideau Street, 9th Floor
Ottawa, Ontario K1A 0K2

CZECH REPUBLIC
DIC Czech Republic-NATO RTO
VTÚL a PVO Praha
Mladoboleslavská ul.
Praha 9, 197 06, Česká republika

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)
Defence Industry & Research General Directorate
Research Directorate
Fakinos Base Camp
S.T.G. 1020
Holargos, 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 di Documentazione Tecnico-Scientifica della Difesa
Via XX Settembre 123a
00187 Roma

LUXEMBOURG
See Belgium

NETHERLANDS
Royal Netherlands Military Academy Library
P.O. Box 90.002
4800 PA Breda

NORWAY
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25, NO-2007 Kjeller

POLAND
Armament Policy 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
Dstl Knowledge Services
Kentigern House, Room 2246
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
Acquisitions
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)

ISBN 92-837-1102-5