REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-03-2011	Technical Paper	MAR 2011 - APR 2011

4. TITLE AND SUBTITLE

Characterizing Deletion Transformations across Dialects using a Sophisticated Tying Mechanism

5a. CONTRACT NUMBER
FA8720-05-C-0002

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Nancy F. Chen, Wade Shen, and Joseph P. Campbell

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

MIT Lincoln Laboratory
244 Wood Street
Lexington, MA 02420

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

NSA
9800 Savage Rd
Ft. Meade, MD 20755

10. SPONSOR/MONITOR'S ACRONYM(S)

NSA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We propose a sophisticated tying mechanism for modeling deletion transformations between dialects. We empirically show that the proposed tying mechanism reduces deletion errors by 33% when compared to a baseline system using a standard tying mechanism. Statistical tests show that the proposed and baseline models make statistically different errors, thus suggesting that they are complementary systems in dialect recognition tasks. Pronunciation rules learned by our proposed system quantify the occurrence frequency of known rules, and suggest rule candidates for further linguistic studies.

15. SUBJECT TERMS

pronunciation model

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	SAR	4	Zach Sweet
U	U	U			19b. TELEPHONE NUMBER *(include area code)* 781-981-5997

MS-56060

Characterizing Deletion Transformations across Dialects using a Sophisticated Tying Mechanism*

Nancy F. Chen[1,2], *Wade Shen*[2], *Joseph P. Campbell*[2]

[1]Massachusetts Institute of Technology, Cambridge, MA, U.S.A.
[2]MIT Lincoln Laboratory, Lexington, MA, U.S.A.
nancyc@mit.edu, swade@ll.mit.edu, jpc@ll.mit.edu

Abstract

We propose a sophisticated tying mechanism for modeling deletion transformations between dialects. We empirically show that the proposed tying mechanism reduces deletion errors by 33% when compared to a baseline system using a standard tying mechanism. Statistical tests show that the proposed and baseline models make statistically different errors, thus suggesting that they are complementary systems in dialect recognition tasks. Pronunciation rules learned by our proposed system quantify the occurrence frequency of known rules, and suggest rule candidates for further linguistic studies.

Index Terms: pronunciation model

1. Introduction

While many dialect recognition systems take advantage of phonotactic differences across dialects, most of these systems do not focus on characterizing linguistically interpretable results. Exceptions include [2, 10, 4, 5]. In [2], acoustic differences caused by phonetic context were used to infer underlying phonetic rules. In [4, 5], where discriminative classiers are trained to recognize dialects, and N-grams or context-dependent phones helpful in dialect recognition are discussed. This line of work has important applications in forensic phonetics [1].

In our previous work [10], we proposed a pronunciation model which characterizes phonetic transformations across dialects. We adopted standard triphone state clustering techniques used in ASR to model context-dependent phonetic transformations across dialects. In this work, we refine our previous model to characterize deletion transformations more appropriately. We show that deletion errors are reduced by 33% compared to our previous standard tying system [10].

2. Method

2.1. Pronunciation Model

We used an HMM system for our previously proposed pronunciation model. The reference dialect's pronunciation is modeled by the states, and the pronunciation of the dialect of interest is modeled by the observations emitted by the states. Phonetic transformations (deletion, insertion, and substitution) between to two dialects are modeled by state transition probabilities.

*This work is sponsored by the Command, Control and Interoperability Division (CID), which is housed within the Department of Homeland Security's Science and Technology Directorate under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

2.1.1. HMM Architecture

States. Suppose the reference phone sequence is $C = c_1, c_2, ..., c_n$. Each reference phone c_i corresponds to two states, a *normal* state s_{2i-1} followed by an *insertion* state s_{2i}. Therefore, the corresponding states of the reference phone sequence C are $S = s_1, s_2, ..., s_{2n}$. $Q = q_1, q_2, ..., q_T$ represents the possible state transition path taken in S. Q takes on values of phones in S by a monotonic order:

$$\text{if} \quad q_t = s_\imath, q_{t+1} = s_\jmath, \quad \text{then} \ \ \imath \leq \jmath. \tag{1}$$

The probability being in state x and emitting observation v_k at time t is

$$B_x(k) = P(o_t = v_k | q_t = x), \tag{2}$$

where $1 \leq x \leq N, 1 \leq k \leq M$.

When traversing over all the possible state transition paths of S, the probability of s_i corresponding to state x and emits v_k is

$$b_i(o_t) = B_x(k), \tag{3}$$

where $s_i = x, 1 \leq i \leq 2n$.

State Transitions. There are 4 types of state transitions: insertion, self-insertion, deletion, and typical transitions. State transition types are represented by $r \in \{ins, sel, del, typ\}$.

The state transition probability from state x to state y through transition arc type r is

$$A_{xry} = P(q_{t+1} = y, r | q_t = x), \tag{4}$$

where $1 \leq x, y \leq N$, transition type $r \in \{ins, sel, del, typ\}$, $\sum_y \sum_r A_{xry} = 1, \forall x$.

When traversing over all the possible state transition paths of S, the probability of transitioning from state s_i to state s_j in S through transition type r is

$$a_{irj} = A_{xry}, \tag{5}$$

where $1 \leq i, j \leq 2n, s_i = x, s_j = y$. Note that if $r = sel$, then $i = j$.

2.1.2. Decision Tree Clustering

In the context of our pronunciation model, a decision tree is grown for each state s, where $1 \leq s \leq N$. At each node k of the tree, a list of attributes are used to split the data into two subgroups. The attribute H_k which generates the *best* split is chosen to split the data to children nodes. This splitting processing is done recursively until a stop criterion is reached. The

best splitting is determined by an objective function such as the log likelihood increase or information gain.

Assume $p_k(j)$ is the probability that state s emits observation v_j at node k, where attribute H_k specifies the subgroups of s that belong to node k. The likelihood function of state s emitting observation v_j at node k is $L(o = v_j | s \in H_k) = p_k(j)$. The maximum likelihood estimate of $p_k(j)$ is simply the observed relative frequency of observation v_j at node k: $\hat{p}_k(j) = \frac{n_k(j)}{n_k}$, where $n_k(j)$ is the expected number of times v_j occurred at node k, and $\sum_j n_k(j) = n_k$. The total likelihood at node k is

$$L(O_k | s \in H_k) = \Pi_{j=1}^{M} p_k(j)^{n_k(j)} \tag{6}$$

Suppose node k_1 and node k_2 are the children of node k, then the log likelihood increase of splitting node k to node k_1 and k_2 is

$$\Delta \log L = log \frac{\Pi_{i=1,2} L(O_{k_i} | s \in H_{k_i})}{L(O_k | s \in H_k)} \tag{7}$$

2.1.3. *Standard Triphone Tying Mechanism*

Standard tying is similar to how triphone states are tied in automated speech recognition [6]. Suppose attribute H_{fk} in the decision tree model corresponds to the feature f being present ($k = 1$) or absent ($k = 2$) of the contextual phones of a triphone state.

The log likelihood of a group of clustered triphone states are computed using the expected number of emissions of these triphones. The expected number of emissions of triphone states q_t that correspond to attribute H_{fk} emitting observation v_j is

$$E(v_j | q_t \in H_{fk}) = \sum_{t=1}^{T} P(O | q_t \in H_{fk}, \lambda, S) \delta(o_t, v_j), \tag{8}$$

where S are the states, and q_t all share the same *center-phone*, $k = \{1, 2\}$, and

$$\delta(\mathbf{o_t}, \mathbf{v_j}) = 1 \text{ if } \mathbf{o_t} = \mathbf{v_j} \tag{9}$$

$$= 0 \text{ otherwise} \tag{10}$$

The total likelihood of $q_t \in H_{fk}$ is

$$L(O | q_t \in H_{fk}) = \Pi_{j=1}^{M} \langle \frac{E(v_j | q_t \in H_{fk})}{\sum_j E(v_j | q_t \in H_{fk})} \rangle^{E(v_j | q_t \in H_{fk})} \tag{11}$$

After state clustering, assume triphone states are clustered into I groups. Group i is specified by $G_i = (\zeta_\ell, \zeta_m, \zeta_r)$, where ζ_ℓ specifies the left context state, ζ_m specifies the center (middle) state, and ζ_r specifies the right context state.

The models estimation equations still have the same form in a typical HMM system [7]:

$$\bar{A_{G_i,r}} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} \in G_i, r, q_t = y | \lambda, S)}{\sum_{t=1}^{T} \sum_r P(\mathbf{O}, q_{t-1} \in G_i | \lambda, S)} \tag{12}$$

$$\bar{B_{G_i}}(k) = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_t \in G_i, | \lambda, S) \delta(o_t, v_k)}{\sum_{t=1}^{T} P(\mathbf{O}, q_t \in G_i, | \lambda, S)} \tag{13}$$

Limitations: The standard triphone tying mechanism makes two assumptions for deletion rules. (1) If a phone is deleted, the pronunciation of its previous phone will be affected and characterized phonetically through automatic phone recognition or manual phone transcriptions. (2) The phone following the deleted phone does not characterize when deletions occur. These assumptions might be over-simplifications and only apply to certain deletion rules. For example, one difference between General American English (GAE) and Received Pronunciation (RP) in British English is that the former is rhotic while the latter is not. Rhotic speakers pronounce /r/ in all positions, while non-rhotic speakers pronounce /r/ only if it is followed by a vowel sound in the same syllable. For instance, the word *park* (/p aa r k/) in American English will sound like *pak* ([p aa: k][1]) in RP, since /r/ is follwed by a consonant /k/. Clearly, this non-rhotic rule does not comply with assumption (2). While the vowel before /r/, /aa/ does changes its vowel quality by becoming longer [aa:], this phenomenon might be too sublet to characterize practically in automated systems, and might not be true for all deletion transformations across dialects.

In addition, since deletions are modeled by deletion transition arcs that skip states (therefore the deleted states will not emit anything) in our model, it is more appropriate to use arc clustering instead of traditional state clustering to determine the tying structure.

2.1.4. *Sophisticated Tying Mechanism*

A state transition arc is specified by the origin state and the destination state. In the case of deletion arcs, the normal states that are skipped during the transition also characterizes the state transition arc.

Consider triphone state $s_{k-1} - s_k - s_{k+1}$. Expected counts of the state x being deleted when q_t corresponds to attribute H_{fk} is

$$E_{d=x} = \sum_{t=1}^{T} \sum_{d=x} P(q_{t+1}, r = del, d | q_t \in H_{fk}), \tag{14}$$

where d represents the deleted state.

$$E_{d \neq x} = \sum_{t=1}^{T} \sum_{x} P(q_{t+1}, r | q_t = x \in H_{fk}), \tag{15}$$

since state x cannot have deleted if there were transition arcs leaving it.

The total likelihood of q_t corresponding to attribute H_{fk} is

$$L(x | q_t \in H_{fk}) = \langle \frac{E_{d=x}}{E_{d=x} + E_{d \neq x}} \rangle^{E_{d=x}} \langle \frac{E_{d \neq x}}{E_{d=x} + E_{d \neq x}} \rangle^{E_{d \neq x}} \tag{16}$$

The log likelihood increase in decision tree clustering can thus be computed to determine the attributes of each group of clustered deletion arcs. After arc clustering, assume deletion arcs are clustered into J groups. Group j is specified by $D_j = (\sigma_j, \varsigma_j, \tau_j)$, where σ_j specifies the source of the arc, ς_j specifies the skipped state, and τ_j specifies the target of the arc. The model estimation equation for deletion transitions belonging to clustered group D_j is similar to Eq. (12):

[1] [aa:] represents a long [aa]

$$A_{D_j} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} \in \sigma_j, r = del, d \in \varsigma_j, q_t \in \tau_j | \lambda, S)}{\sum_{t=1}^{T} \sum_r P(\mathbf{O}, q_{t-1} \in \sigma_j | \lambda, S)} \quad (17)$$

After state clustering, assume triphone states are clustered into I groups. Group i is specified by $G_i = (\zeta_\ell, \zeta_m, \zeta_r)$, where ζ_ℓ specifies the left context state, ζ_m specifies the center (middle) state, and ζ_r specifies the right context state.

After arc clustering, assume deletion arcs are clustered into J groups. Group j is specified by $D_j = (\sigma_j, \varsigma_j, \tau_j)$, where σ_j specifies the source of the arc, ς_j specifies the skipped state, and τ_j specifies the target of the arc.

Suppose we want to compute the tied probabilities of the the triphone state $s_{k-1} - s_k - s_{k+1}$, where $s_{k-1} \in \zeta_\ell$, $s_k \in \zeta_m$, and $s_{k+1} \in \zeta_r$. We first compute all the clustered deletion probabilities originating from s_k. Then we estimate the typical and insertion transition probabilities as in the standard tying case using a new Lagrange constraint.

The sum of all deletion probability leaving triphone state $s_{k-1} - s_k - s_{k+1}$ is

$$P_D = P(q_{t+1} = s_{k+2}, r = del | q_t = s_k) \quad (18)$$

$$= \sum_j P(q_{t+1} = s_{k+2} \in \tau_j, r = del | q_t = s_k, s_{k+1} \in \varsigma_j) \quad (19)$$

$$P(s_{k+1} \in \varsigma_j, q_{t+1} = s_{k+2}, r = del | q_t = s_k) \quad (20)$$

where $P(q_{t+1} = s_{k+2} \in \tau_j, r = del | q_t = s_k, s_{k+1} \in \varsigma_j)$ was already computed in Eq. (17) as A_{D_j}.

$$A_{G_i,r} = \frac{\sum_{t=1}^{T} P(\mathbf{O}, q_{t-1} \in G_i, r | \lambda, S)}{\sum_{t=1}^{T} \sum_r P(\mathbf{O}, q_{t-1} \in G_i \cap \Sigma^c | \lambda, S)} (1 - P_D) \quad (21)$$

where $\Sigma = \{\sigma_1 \cup \sigma_2 \cup ... \sigma_I\}$, and $\Sigma^c = \{\sigma_1^c \cap \sigma_2^c \cap ... \sigma_I^c\}$, $r \in \{typ, ins\}$.

2.2. Statistical Test

We used the matched pairs test in [9] to evaluate whether the performance difference of the two systems being compared are statistically significant.

Let us suppose that we can divide the output stream from a pronunciation model system into segments in such a way that the errors in one segment are statistically independent of the errors in any other segment. Suppose we are comparing the performance difference of S_1 and S_2 Let N_1^i be the number of errors made on the i i-th segment by System S_1 , and N_2^i the number of errors made by System S_2. Note that the type of error is unimportant, as long as the method of counting errors is consistent for each segment and for both systems.

Let $Z_i = N_1^i - N_2^i$, $i = 1, ..., n$, where n is the number of segments. Let μ_z be the unknown average difference in the number of errors in a segment made by the two Systems. We would like to ascertain whether $\mu_z = 0$. The maximum likelihood estimate of μ_z and the variance of Z_i are

$$\hat{\mu_z} = \sum_{i=1}^{n} \frac{Z_i}{n} \quad (22)$$

$$\hat{\sigma_z^2} = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \mu_z)^2 \quad (23)$$

Table 1: WSJCAM0 data partition

Set	Speaker number	Duration
Train	92	15.3 hr
Dev	48	4 hr
Test	48	4 hr

If W is defined as

$$W = \frac{\hat{\mu_z}}{\hat{\sigma_z}/\sqrt{n}}, \quad (24)$$

then if n is large enough, W will approximate a standard normal distribution $N(0, 1)$. We can test the null hypothesis $H0$: $\mu_z = 0$, bu computing $P = 2Pr(Z > |w|)$, where Z is a random variable with distribution $N(0, 1)$ and w is the realized value of W.

3. Experiments

3.1. Assumptions and Protocol

We adapt the assumptions in [10] to the following.

1. All pronunciation variations across dialects are governed by underlying phonetic rules.
2. The ground-truth surface phones of the WSJCAM0 corpus are the phonetic transcriptions it provides.
3. The ability to predict ground-truth surface phones using the trained pronunciation models indicates how well the underlying phonetic rules are retrieved from the pronunciation model algorithms.

3.2. Data

The speech database used is WSJ-CAM0 is the UK English equivalent of a subset of the US American English WSJ0 database [11]. The data partition of WSJ-CAM0 is listed in Table 1.

3.3. Implementation Details

3.3.1. Pronunciation Model

Given the trained pronunciation model, we generate the most likely observations given the reference phones, and compare the generated observations with the ground-truth observations, provided by the phone transcriptions in WSJCAM0. Reference phones are determined by the American English dictionary given the text.

3.3.2. Statistical Test

We could divided the generated surface phone outputs into segments where no errors have occurred for some minimal time period T ("good" segments) and segments where errors occur ("bad" segments), according to [9]. T is required to be sufficiently long to ensure that after a good segment, the rst error in a bad segment is independent of any previous errors. T was swept on the development set (ranging from values of 9 to 402), and all resulted in similar p-values ($p \ll 0.001$) on the test set. The number of segments n ranged from 756 to 32491, which is assumed to be sufficiently large enough for W to be normally distributed, and a good estimate of the variance of Z_i can be obtained. Errors were divided into deletion, insertion, and substitution, and each type of error was analyzed separately.

3.4. Phone Error Rate Results

The phone error rate (PER) between the ground-truth surface phones and the generated surface phone of each system are

Table 2: Phone error rate (PER) for each system. Units are in %. Total number of phones in the test set: 299,853.

System	Overall	Deletion	Insertion	Substitution
Monophone	15.1	2.0	3.3	9.8
Standard Tying	9.0	2.1	1.9	5.0
Sophisticated Tying	9.0	1.4	2.6	5.0

Table 3: Relative PER improvement (compared to baseline Monophone System). Units are in %.

System	Overall	Deletion	Insertion	Substitution
Standard Tying	40	-5	42	49
Sophisticated Tying	40	30	18	49

listed in Table 2. The sub-error categories of deletion, insertion, and substitution are also listed in Table 2. The relative improvement of all systems compared to the baseline monophone system is listed in Table 3. All improvements are shown to be statistically significant ($p \ll 0.0001$) according to the matched pairs test described in Section 2.2.

4. Discussion

4.1. Context-Dependent Systems vs. Monophone System

All systems that exploits context information outperformed the baseline monophone system by 40 % relative ($p \ll 0.001$.). These results verify that phonetic context information is important in characterizing dialect differences, as reported in [10, ?].

If we break down the PER into sub-error categories of insertion, deletion, and substitution, we see statistically significant improvements for all categories in all systems, except for deletion errors for the standard tying system. Compared to the baseline monophone system, the standard tying systems show statically significant negative improvement in deletion errors (-5 %; $p \ll 0.0001$). This result imply that the standard tying systems are over-generalizing deletion rules.

Note that in the standard tying systems, the phone following the deleted phone is never used to characterized the deletion transitions. In the monophone system, deletion rules are characterized by the phone preceding the deleted phone. Phonologically speaking, the phone of interest is generally influenced more by its following phone than its preceding phone. In non-rhotic dialects of English, we also know that the right-context of /r/ is more important in specifying non-rhoticity than the left. Therefore, without characterizing deletion transitions using the right-context phone, it is expected that deletion rules are over-generalized. We expect that including the phone following the deleted phone could characterize deletion transitions more accurately. We discuss these details in the next section.

4.2. Standard Tying vs. Sophisticated Tying

4.2.1. deletion errors

The overall PER between the standard and sophisticated tying systems are the same. The matched pairs test shows that the two systems ae making statistically different errors ($p \ll 0.0001$). If we consider deletion errors, we see that the sophisticated tying system beats the standard tying system by 33% relative. Among the /r/'s that were incorrectly deleted in the standard tying system, the sophisticated tying system correctly generated 24% of these /r/'s. This result also supports the hypothesis that r-dropping rules are characterized by the right context of /r/.

4.2.2. Arc clustering vs. state clustering

In terms of generating dialect-specific pronunciations, standard and sophisticated tying might not show statistical difference, but arc clustering is much more suitable in discovering and interpreting deletion rules. The arc clustering scheme explicitly characterize deletion rules: the decision tree clustering results show potential deletion rule candidates. On the other hand, it is much more challenging to linguistically characterize deletion transformation as phonetic rules in state clustering. Therefore, depending on the need of the task, different tying schemes might be preferred. For generating dialect-specific pronunciations or dialect recognition tasks, state clustering is simpler to implement and still performs well. If speech science analysis is required, arc clustering is more suitable in characterizing deletion rules.

4.3. Implications for Dialect Recognition

The statistical test evaluates whether two systems make the unique errors. The signicant statistical test results indicate that the two systems being compared makes different errors, implying that if the pronunciation models are used in dialect recognition tasks, they will fuse well.

5. Conclusions

We propose a sophisticated tying mechanism for modeling deletion transformations between dialects. We empirically show that the proposed tying mechanism reduces deletion errors by 33% when compared to a baseline system using a standard tying mechanism. Statistical tests show that the proposed and baseline models make statistically different errors, thus suggesting that they are complementary systems in dialect recognition tasks. Pronunciation rules learned by our proposed system quantify the occurrence frequency of known rules, and suggest rule candidates for further linguistic studies. Potential applications include forensic phonetics, accent training, and dialect recognition.

6. References

[1] Rose, P., "Forensic Speaker Identification, Taylor and Francis, 2002.

[2] Chen, N., Shen, W., Campbell, J., "A Linguistically-Informative Approach to Dialect Recognition Using Dialect-Specific Context-Dependent Phonetic Models," ICASSP, 2010.

[4] Biadsy, F. *et al.*,"Discriminative phonotactics for dialect recognition using context-dependent phone classifiers," Odyssey, 2010.

[5] Richardson, F. *et al.*, "Discriminative N-Gram Selection for Dialect Recognition," Interspeech, 2009.

[6] Fosler-Lussier, E., "A Tutorial on Pronunciation Modeling for LVSR," in Renals, S. and Grefenstette, G., editors, Text and Speech Triggered Info, Springer-Verlag Berlin Heidelberg, 2003.

[7] Rabiner, L.,Juang, B. H., "Fundamentals of Speech Recognition," Prentice Hall, 1993.

[8] R. Quinlan, "Induction of decision trees, Machine Learning, Vol. 1, No. 1, pp.81-106, 1986.

[9] Gillick, L., Cox, S., "Some statistical issues in the comparison of speech recognition algorithms," Proc. ICASSP, 1989.

[10] Chen, N. *et al.*, "Informative Dialect Recognition using Context-Dependent Pronunciation Modeling,", to appear in Proc. ICASSP, 2011.

[11] Robinson, T. *et al.*, "WSJ-CAM0: A British English Corpus for Large Vocabulary Continuous Speech Recognition", ICASSP, 1994.