# Characterizing Deletion Transformations across Dialects using a Sophisticated Tying Mechanism

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

## 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.

## Subject Terms
- pronunciation model

## Security Classification
- **U**

## Number of Pages
4

---

15. **REPORT DOCUMENTATION PAGE**

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

<table>
<thead>
<tr>
<th>1. REPORT DATE (DD-MM-YYYY)</th>
<th>2. REPORT TYPE</th>
<th>3. DATES COVERED (From - To)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30-03-2011</td>
<td>Technical Paper</td>
<td>MAR 2011 - APR 2011</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4. TITLE AND SUBTITLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characterizing Deletion Transformations across Dialects using a Sophisticated Tying Mechanism</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5a. CONTRACT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>FA8720-05-C-0002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5b. GRANT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5c. PROGRAM ELEMENT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6. AUTHOR(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nancy F. Chen, Wade Shen, and Joseph P. Campbell</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT Lincoln Laboratory</td>
</tr>
<tr>
<td>244 Wood Street</td>
</tr>
<tr>
<td>Lexington, MA 02420</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>8. PERFORMING ORGANIZATION REPORT NUMBER</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSA</td>
</tr>
<tr>
<td>9800 Savage Rd</td>
</tr>
<tr>
<td>Ft. Meade, MD 20755</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>10. SPONSOR/MONITOR’S ACRONYM(S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>11. SPONSOR/MONITOR’S REPORT NUMBER(S)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>12. DISTRIBUTION / AVAILABILITY STATEMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.</td>
</tr>
</tbody>
</table>

| 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

<table>
<thead>
<tr>
<th>16. SECURITY CLASSIFICATION OF:</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>a. REPORT b. ABSTRACT c. THIS PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>17. LIMITATION OF ABSTRACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>18. NUMBER OF PAGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>19a. NAME OF RESPONSIBLE PERSON</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zach Sweet</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>19b. TELEPHONE NUMBER (include area code)</th>
</tr>
</thead>
<tbody>
<tr>
<td>781-981-5997</td>
</tr>
</tbody>
</table>

---

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18
Characterizing Deletion Transformations across Dialects using a Sophisticated Tying Mechanism*

Nancy F. Chen\textsuperscript{1,2}, Wade Shen\textsuperscript{2}, Joseph P. Campbell\textsuperscript{2}

\textsuperscript{1}Massachusetts Institute of Technology, Cambridge, MA, U.S.A.; \textsuperscript{2}MIT Lincoln Laboratory, Lexington, MA, U.S.A.

nancyc@mit.edu, swadesll.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 \cite{2,10,4,5}. In \cite{2}, acoustic differences caused by phonetic context were used to infer underlying phonetic rules. In \cite{4,5}, where discriminative classifiers 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 \cite{1}.

In our previous work \cite{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 \cite{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 (C\&ID), 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.
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 
\[
L(O_s | s \in H_k) = \prod_{t=1}^{T} p_k(j)^{n(k,j)}
\]

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
\[
\Delta \log L = \log \frac{\prod_{t=1}^{T} p_k(j)^{n(k,j)}}{\prod_{t=1}^{T} p_{k_1}(j)^{n(k_1,j)}p_{k_2}(j)^{n(k_2,j)}}
\]

2.1.3. Standard Triphone Tying Mechanism

Standard tying is similar to how triphone states are tied in automated speech recognition [6]. Suppose attribute \( H_{jk} \) 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 \( H_{jk} \) emitting observation \( v_j \) is
\[
E(v_j | q_t \in H_{jk}) = \sum_{t=1}^{T} P(O_t | q_t \in H_{jk}, x) \delta(o_t, v_j)
\]

where \( S \) are the states, and \( q_t \) all share the same center-phone, \( k = \{1, 2\} \), and
\[
\delta(o_t, v_j) = \begin{cases} 1 & \text{if } o_t = v_j \\ 0 & \text{otherwise} \end{cases}
\]

The total likelihood of \( q_t \in H_{jk} \) is
\[
L(O | q_t \in H_{jk}) = \prod_{j=1}^{I} \frac{E(v_j | q_t \in H_{jk}) \delta(o_t, v_j)}{E(v_j | q_t \in H_{jk})}
\]

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

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

\[
A_{G_i,r} = \frac{\sum_{t=1}^{T} P(O_t, q_{t-1} \in G_i, r; q_t = y | \lambda, S)}{\sum_{t=1}^{T} P(O_t, q_{t-1} \in G_i | \lambda, S)}
\]

\[
B_{G_i}(k) = \frac{\sum_{t=1}^{T} P(O_t, q_t \in G_i | \lambda, S) \delta(o_t, v_k)}{\sum_{t=1}^{T} P(O_t, q_t \in G_i | \lambda, S)}
\]

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 a rr k) in American English will sound like pak (p a k)\(^1\) in RP, since /r/ is followed by a consonant /k/. Clearly, this non-rhotic rule does not comply with assumption (2).

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 \( H_{jk} \) is
\[
E_{d_{x}x} = \sum_{t=1}^{T} P(q_{t+1}, r = del, d | q_t \in H_{jk})
\]

where \( d \) represents the deleted state.

\[
E_{d_{x}x} = \sum_{t=1}^{T} P(q_{t+1}, r | q_t = x \in H_{jk})
\]

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

The total likelihood of \( q_t \) corresponding to attribute \( H_{jk} \) is
\[
L(x | q_t \in H_{jk}) = \frac{E_{d_{x}x}}{E_{d_{x}x} + E_{d_{x}a}}
\]

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 \( D_j = (\sigma_j, c_j, r_j) \), where \( \sigma_j \) specifies the source of the arc, \( c_j \) specifies the skipped state, and \( r_j \) specifies the target of the arc. The model estimation equation for deletion transitions belonging to clustered group \( D_j \) is similar to Eq. (12):
After state clustering, assume triphone states are clustered into \( J \) groups. Group \( j \) is specified by \( C_j = (\xi_j, \zeta_m, \zeta_r) \), where \( \xi_j \) specifies the left context state, \( \zeta_m \) specifies the center (middle) state, and \( \zeta_r \) specifies the target context state.

After arc clustering, assume deletion arcs are clustered into \( J \) groups. Group \( j \) is specified by \( D_j = (\sigma_j, \varepsilon_j, \tau_j) \), where \( \sigma_j \) specifies the source of the arc, \( \varepsilon_j \) specifies the skipped state, and \( \tau_j \) 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 \( s_{k-1} \in C_k \), \( s_k \in \zeta_m \), 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

\[
P_D = P(q_{k+1} = s_{k+2}, r = \text{del}|q_k = s_k) \]

\[
= \sum_j P(q_{k+1} = s_{k+2} \in \tau_j, r = \text{del}|q_k = s_k, s_{k+1} \in \varepsilon_j) \]

\[
P(s_{k+1} \in \varepsilon_j, q_{k+1} = s_{k+2}, r = \text{del}|q_k = s_k) \]

where \( P(q_{k+1} = s_{k+2} \in \tau_j, r = \text{del}|q_k = s_k, s_{k+1} \in \varepsilon_j) \) was already computed in Eq. (17) as \( A_{Dj} \).

\[
A_{G_{ij}r} = \frac{\sum_{\ell=1}^{J} \sum_r P(O, q_{k-1} \in G_j, r|\lambda, S)}{\sum_{\ell=1}^{J} \sum_r P(O, q_{k-1} \in C_j \cap \Sigma^c_l|\lambda, S)} (1 - R_{k-1}) \]

\[
\sigma = \{ s_{k-1} \cup s_k \cup \ldots \sigma_{ij} \}, \quad \Sigma^c = \{ \sigma^c_1 \cap \sigma^c_2 \cap \ldots \}.
\]

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 \( N_j \) be the number of errors made on the \( i \)-th segment by System \( S_1 \), and \( N_j^2 \) 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 \( Z_i = N_j^1 - N_j^2, i = 1, \ldots, n \), where \( n \) is the number of segments. Let \( \mu_n \) be the unknown average difference in the number of errors in a segment made by the two Systems. We would like to ascertain whether \( \mu_n = 0 \). The maximum likelihood estimate of \( \mu_n \) and the variance of \( Z_i \) are

\[
\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i \]

\[
\hat{\sigma}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Z_i - \hat{\mu}_n)^2
\]

If \( W \) is defined as

\[
W = \frac{\mu_n}{\hat{\sigma}_n^2 / \sqrt{n}},
\]

then if \( n \) is large enough, \( W \) will approximate a standard normal distribution \( N(0,1) \). We can test the null hypothesis \( H_0: \mu_n = 0 \), by computing \( P = 2P(Z > |w|) \), 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 divide 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 \( (p < 0.001) \) on the test set. The number of segments \( n \) ranged from 750 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
24% of the system's results also support the hypothesis that systems are making statistically significant negative improvements in 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 < 0.0001). This result implies 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 characterize 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 by 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 are making statistically different errors (p < 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. 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 characterizes 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 significant 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