Characterizing Phonetic Transformations and Acoustic Differences Across English Dialects*

Nancy F. Chen, *Member, IEEE,* Sharon W. Tam, *Member, IEEE,* Wade Shen, *Senior Member, IEEE,* and Joseph P. Campbell, *Fellow, IEEE*

Abstract— **In this work, we propose a framework that automatically discovers dialect-specific phonetic rules. These rules characterize when certain phonetic or acoustic transformations occur across dialects. To explicitly characterize these dialect-specific rules, we adapt the conventional hidden Markov model to handle insertion and deletion transformations. The proposed framework is able to convert pronunciation of one dialect to another using learned rules, recognize dialects using learned rules, retrieve dialect-specific regions, and refine linguistic rules. Potential applications of our proposed framework include computer-assisted language learning, sociolinguistics, and diagnosis tools for phonological disorders.**

Index Terms—**accent, phonological rules, informative dialect recognition, phonotactic modeling, pronunciation model**

I. Introduction

Dialect is an important, yet complicated, aspect of speaker variability. In this work, we define dialects to be sub-varieties of the same language where these sub-varieties are mutually intelligible and their writing systems are convertible if not the same (e.g., British English and American English; General American English and African American Vernacular English; Mandarin spoken in China, Singapore, Taiwan).[1] While dialect differences arise from every level of the linguistic hierarchy (e.g., acoustic, phonetic, vocabulary, syntax, prosody [1]), we focus on analyzing dialect differences at the phonetic and acoustic levels.[2]

Dialect differences (at the phonetic and acoustic levels) are often attributed as phonetic rules by sociolinguists through manual analyses [3]. A phonetic rule specifies the phonetic context of when a phone transforms to another phonetic or acoustic identity [4], [5]. For example, the R-dropping rule of the received pronunciation dialect in British English [6]: [r] $\rightarrow \emptyset /$ __ [+consonant], specifies that the phone [r] is deleted ($\emptyset$) when it is followed by a consonant.

The motivation of our work is to automate the explicit characterization of such phonetic rules, which currently relies heavily on manual analysis in fields like sociolinguistics. The time-consuming nature of manual analyses often limits the amount of data analyzed, which could potentially compromise the validity of these phonetic rules. Although researchers [7] have advocated using automatic procedures to increase efficiency and effectiveness in phonetic analysis, automatic processing remains a rare practice in the sociolinguistic community.

In addition to discovering scientific truth in sociolinguistics, there are various practical applications for automating the characterization of pronunciation patterns. For example, in computer-assisted language learning, it is desirable to pinpoint non-native pronunciation patterns to help users acquire a second language; in speech pathology, clinicians benefit from automatic characterizations of atypical articulatory or phonological patterns to diagnose and treat patients; and in forensic phonetics, it is necessary that results of a dialect recognizer are justifiable on linguistic grounds [8].

In this work, we focus on automatically analyzing dialect differences in the form of phonetic rules. To this end, we design an automatic system to discover possible phonetic rules, to quantify how well these rules discriminate dialects, and to provide human analysts (e.g., forensic phoneticians) with regions-of-interest for further analysis and verification.

A. Background on Automatic Dialect Recognition

Automatic dialect recognition has historically been cast as a language recognition problem. Thus, mainstream dialect recognition approaches are ported from language recognition [9], focusing on the acoustic level (e.g., [15], [13], [14], [10], [11], [12], [17]) or phonetic level (e.g., [16], [21], [18], [19], [20]).

Typical acoustic approaches model cepstral features, such as mel-frequency cepstral coefficients (MFCC) and shifted delta cepstrum (SDC) [22]. Common model choices include Gaussian mixture models (GMM). While GMMs achieve good performance, they do not provide insight into where dialect differences occur. Adapted phonetic models [13] are an extension of GMM, where acoustic information is modeled in phonetic categories, making it easier to pinpoint where the acoustic differences lie. In our previous work [15], acoustic differences caused by phonetic context were further used to

N. F. Chen was with MIT Lincoln Laboratory, USA. She is now with Institute for Infocomm Research, Singapore. Email: nancychen@alum.mit.edu.

S. W. Tam is with MIT Lincoln Laboratory, USA. Email: sharon.tam@ll.mit.edu.

W. Shen is with MIT Lincoln Laboratory, USA. Email: swade@ll.mit.edu.

J. P. Campbell is with MIT Lincoln Laboratory, USA. Email: j.campbell@ieee.org.

*This work is sponsored by the Department of Defense 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.

[1] Note that language varieties such as Mandarin and Cantonese are often colloquially referred to as Chinese dialects. However, from the linguistic viewpoint, Cantonese and Mandarin are two different Chinese languages [2], since they have different phonological, tonal, and syntactic systems.

[2] Dialect differences at the phonetic and acoustic levels are also referred to or perceived as *accents*.

infer underlying phonetic rules, making the dialect recognizer more informative to humans.

Phonetic approaches exploit phonotactic differences across dialects. One classic system is Phone Recognition followed by Language Modeling (PRLM) [9]. It exploits a larger range of phonological differences than GMM by modeling N-grams of decoded phone sequences. For example, the bigram $\{$[aa] [r]$\}$, found in words like park, heart, rarely occurs in British English [6], while it is more common in American English. This phonotactic difference is modeled implicitly[3] in PRLM. Unlike PRLM, in this work, we characterize these dialect differences *explicitly* as rules, which are more human interpretable.

B. Proposed Framework for Automatically Analyzing Dialect-Specific Rules

This paper extends our previously published conference articles [16], [23], [24] and contains detailed experiments, analyses, and discussions left out in the conference versions. Below we sketch out the proposed model and the evaluation framework.

1) Model: We adopt the concept of pronunciation modeling [25] from automatic speech recognition (ASR) to analyze dialects. We adapt the traditional HMM to explicitly model phonetic transformations: substitutions, deletions, and insertions. For example, the R-dropping rule [6] is a type of deletion transformation (American: [p aa r k] ⇒ British: [p aa k]). Fig. 1 shows a comparison between the traditional and proposed HMM frameworks. Our proposed framework provides flexibility in characterizing phonetic transformations and can be easily refined to accommodate particular considerations of interest. For example, to learn more refined deletion rules, we propose an arc clustering scheme for deletion transitions to determine the tying structure used during parameter estimation in Section II-D.

2) Evaluation and Analysis: We evaluate two pairs of English dialects. The first pair is American and British English. We conduct a pronunciation conversion experiment in Section III, where we compare the different variants of the proposed phonetic pronunciation models. In this experiment, it is assumed that if a model is able to convert American English pronunciation to British English, the model has learned the phonetic rules governing the dialect differences well.

The second pair of dialects is African American Vernacular English (AAVE) and non-AAVE American English. We break down the experiments into two parts. The first part quantifies how well the learned rules explain dialect differences via dialect recognition tasks. The second part assesses how well dialect-specific rules are characterized via information retrieval tasks. In addition, we further analyze how the proposed system might complement traditional linguistic analyses of dialect-specific rules.

The rest of the paper is organized as follows. In Section II, we present the mathematical framework to explicitly model phonetic and acoustic transformations across dialects and refine the model to more appropriately characterize deletions. We also discuss the relationship of our framework with other systems. In Section III, we evaluate how well the learned phonetic rules are able to convert American English pronunciation to British English. In Section IV, we use dialect recognition performance to quantify the dialect differences characterized by learned rules. In Section V, we examine how well the proposed framework retrieves dialect-specific regions defined by linguistic rules and discuss how our framework helps phoneticians analyze rules. In Section VI, we discuss the limitations of the model assumptions and possible applications of the proposed framework. Finally, we conclude our work in Section VII.

II. Proposed Model

In Section II-A, we describe the input to the proposed model and our aim. In Section II-B, we illustrate why the network of traditional HMMs is not suitable in modeling dialect-specific rules. In Section II-C, we describe how we alter the traditional HMM network to explicitly characterize phonetic transformations by further categorizing state transitions and introducing insertion states. We also present a variant model where arc clustering is used to refine deletion rules in Section II-D. In Section II-E, we show that our proposed framework can be extended to model acoustic differences. We compare our models with other dialect recognition systems in Section II-F.

A. Input and Aim of Model

Suppose we are given a dataset consisting of speech utterances (e.g., British English), their corresponding word transcriptions and a reference pronunciation dictionary (e.g., American English). For each speech utterance, we can generate reference phones $C = c_1, c_2, ..., c_n$ (obtained using the word transcriptions and the reference pronunciation dictionary) and surface phones $O = o_1, o_2, ..., o_T$ (which can be obtained through either automatic or manual phonetic transcription, depending on the available resources and the experimental design).

We want to model the relationship between the reference phones (e.g., American English pronunciation) and the surface phones (e.g., British English pronunciation). Such a model will inform us when and where these two dialects differ and how often these dialect differences occur. These dialect differences can be formulated into pronunciation rules, which can be used to automatically recognize dialects and help phoneticians discover new rules or refine/verify existing rules.

B. Traditional HMM

In Fig. 1 (a), a traditional HMM network is shown: reference phones are modeled by states (denoted as circles) and the surface phones are modeled by observations (denoted as squares); the emission of an observation from a state is denoted by a dotted line. An example is shown where the word is *part*, reference phones are [p aa r t], and surface phones are [p ih aa t]. This example illustrates why the traditional HMM

[3] Errors in phone decoding could disguise the underlying bigram ([aa] [r]) to appear in different forms. However, this phonotactic difference is still modeled as long as the phone recognition errors are consistent and not random.

network is not suitable for modeling insertions[4] and deletions[5]: the inserted [ih] surface phone has no corresponding state, and the deleted [r] reference phone has no corresponding observation; these non-correspondences are represented by the green question marks.

C. Phonetic Pronunciation Model (PPM)

In contrast, the network of PPM is designed to model deletions and insertions.

1) States: Suppose we are given a reference phone sequence $C = c_1, c_2, ..., c_n$. Each reference phone c_i corresponds to two states, a *normal* state s_{2i-1} followed by an *insertion* state s_{2i}. The corresponding states of the reference phone sequence C are $S = s_1, s_2, ..., s_{2n}$.[6]

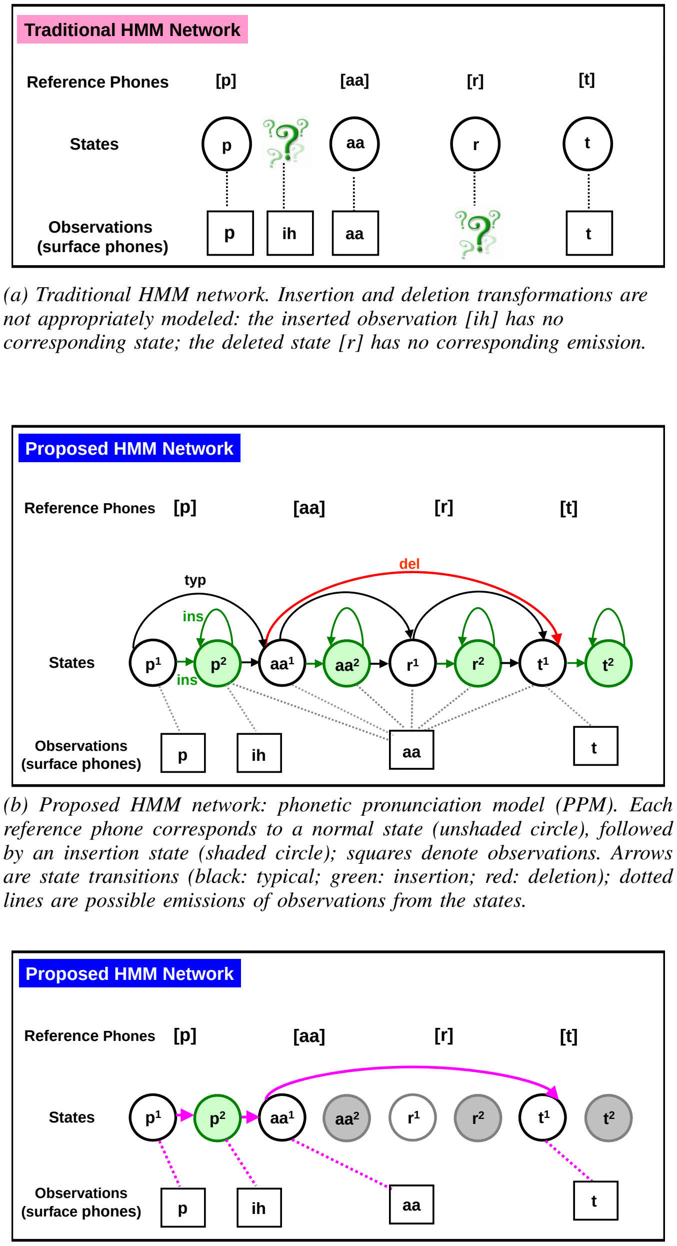
In Fig. 1 (b), normal states are denoted by unshaded circles and insertion states are denoted by shaded circles. The reference phone sequence $C = (c_1, c_2, c_3, c_4) =$ ([p], [aa], [r], [t]) corresponds to the state sequence $S = (s_1, s_2, s_3, s_4, s_5, s_6, s_7, s_8) = (\mathrm{p}^1, \mathrm{p}^2, \mathrm{aa}^1, \mathrm{aa}^2, \mathrm{r}^1, \mathrm{r}^2, \mathrm{t}^1, \mathrm{t}^2)$.

2) Observations: $V = \{v_1, ..., v_M\}$ is the observation alphabet; M is the total number of distinctive phones. The given observation sequence $O = o_1, o_2, ..., o_T$ is illustrated as squares in Fig. 1 (b), where T is the number of observations (surface phones). Note that in general the length of the states and observation length are different; i.e., $2n \neq T$.

The emission of an observation from a state is denoted by dotted lines connecting a circle to a square.

3) Alignment between States and Observations: We define the auxiliary state sequence $Q = q_1, q_2, ..., q_T$. Q takes on values in the state sequence S by a monotonic order: if $q_t = s_i, q_{t+1} = s_j$, then $i \leq j$. Q can be viewed as a means to *align* the state sequence S to the observation sequence O, so that each observation o_i corresponds to a state q_i, where $1 \leq i \leq T$. An alignment example is shown in Fig. 1 (c), where $Q = (q_1, q_2, q_3, q_4) = (s_1, s_2, s_3, s_7) = (\mathrm{p}^1, \mathrm{p}^2, \mathrm{aa}^1, \mathrm{t}^1)$ and emits the observations $O = (o_1, o_2, o_3, o_4) = ([p], [ih], [aa], [t])$.

4) State Transitions: There are 3 types of phonetic transformations: deletion, insertion, and substitution, which are modeled by 3 types of state transitions: deletion, insertion, and typical, respectively. A deletion state transition[7] is defined as a state transition skipping any number of normal states. An insertion state transition is defined as a state transition originating from a normal state and arriving at its insertion state or a state transition that originates and returns to the same insertion state (thus allowing consecutive insertions). All other state transitions are considered typical. We denote state transitions as $r \in \{del, ins, typ\}$. In Fig. 1 (b), the colors of the arcs denote the different types of state transitions (red: deletion; green: insertion; black: typical).



(a) Traditional HMM network. Insertion and deletion transformations are not appropriately modeled: the inserted observation [ih] has no corresponding state; the deleted state [r] has no corresponding emission.

(b) Proposed HMM network: phonetic pronunciation model (PPM). Each reference phone corresponds to a normal state (unshaded circle), followed by an insertion state (shaded circle); squares denote observations. Arrows are state transitions (black: typical; green: insertion; red: deletion); dotted lines are possible emissions of observations from the states.

(c) A possible alignment path in PPM where $Q = (q_1, q_2, q_3, q_4) = (s_1, s_2, s_3, s_7) = (p^1, p^2, aa^1, t^1)$ *and emits observations* $O = (o_1, o_2, o_3, o_4) = ([p], [ih], [aa], [t])$.

Fig. 1. *Comparison between traditional and proposed HMM networks.*

In Fig. 1 (c) the state transitions taken in the alignment example are denoted by pink arcs. Note that for a given $q_t = s_i, q_{t+1} = s_j$, the arc r joining s_i and s_j can be inferred. For example, in Fig. 1 (c), $q_3 = s_3$ and $q_4 = s_7$, implying that the arc r joining s_3 and s_7 is a deletion arc.

5) Model Parameters: The model parameters are the state transition probability and the emission probability. The state transition probability from state x to state y through transition arc type r is

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

where $1 \leq x, y \leq N$, $\sum_y \sum_r A_{xry} = 1, \forall x$.

[4] An *insertion* is defined as a surface phone with no corresponding reference phone.

[5] A *deletion* is defined as a reference phone with no corresponding surface phone.

[6] In actuality, there are two additional special states s_0 and s_{2n+1} that allow for the states s_1 and s_{2n} to be deleted.

[7] While in principle we can allow for skipping multiple normal states in a deletion state transition, in this work we only consider a single normal state to be skipped. In practice, we find that skipping multiple normal states consecutively does not occur often enough to reliably estimate its probability.

The probability of emitting observation v_k at any time t given state x is

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

where $1 \leq x \leq N, 1 \leq k \leq M$. The pronunciation model is denoted as $\lambda = \{A, B\}$. The state transition probability A and emission probability B can be derived similarly to traditional HMM systems [26] by using the Baum-Welch algorithm. The likelihood of the observations given the model $P(O|\lambda)$ can be computed using the forward and backward algorithms to sum up all possible auxiliary state sequence Q.

6) Decision Tree Clustering: Phonetic context is important both for recognizing dialects and understanding their differences, so we want to consider triphones instead of just monophones. However, because triphone modeling requires more parameters, we use decision tree clustering [27] to tie triphone states.

Assume we are given an initial HMM model. Consider a reference (mono)phone modeled by the state x, which emits observations O_k. The state x can be split into two subgroups using attribute H_f: $x \in H_f$ and $x \notin H_f$, which emit observations O_{k_1} and O_{k_2}, respectively; $O_{k_1} \cup O_{k_2} = O_k$; attribute H_f specifies the triphone context of state x. The log likelihood increase of this split is

$$\Delta \log L = \log \frac{L(O_{k_1} | x \in H_f) L(O_{k_2} | x \notin H_f)}{L(O_k | x)}, \tag{3}$$

where L represents likelihood. The attribute chosen to split x is $\arg\max_{H_f} \Delta \log_L$; i.e., the attribute that provides the most likelihood increase. Since the numerator is always greater than the denominator in Eq. (3), this splitting procedure is done recursively until a stop criterion is reached.

For example, consider the phone [r] and the attribute that makes the likelihood increase the most after splitting is $\hat{H_f}$. In the R-dropping case of British English, $\hat{H_f}$ is likely to specify that the phone following [r] is a consonant, splitting the state x into two sub-groups: one representing [r] followed by consonants and the other group representing [r] not followed by consonants.

D. Model Extension: Arc Clustering for Deletions

1) Motivation: Triphone state clustering used in the previous section makes two implicit assumptions about deletion transformations: (1) The phone preceding a deleted phone changes its sound quality; (2) The phone following a deleted phone does not specify when deletions occur. However, these constraints do not always apply. For example, rhotic English (e.g., General American English) pronounces /r/ in all positions, while non-rhotic English (e.g., Received Pronunciation in UK) only pronounces /r/ when it is followed by a vowel [6]. Therefore the word *part* is pronounced as [p aa r t] for an American speaker, but [p aa: t][8] for a British speaker. When modeling the deletion probability through triphone state clustering, the triphone state representing [p-aa+r][9] does not include the right context of the deleted [r] (i.e., [t], a consonant), even though the right context of [r] is crucial in specifying the deletion rule. Recall that deletion transition arcs originate from some state prior to the deleted state and arrives at some state after the deleted state. Therefore, in the previous example, it might make more sense to cluster all the deletion arcs that skip [r] and arrive at a consonant (e.g., [t]), instead of clustering triphone states such as [p-aa+r].

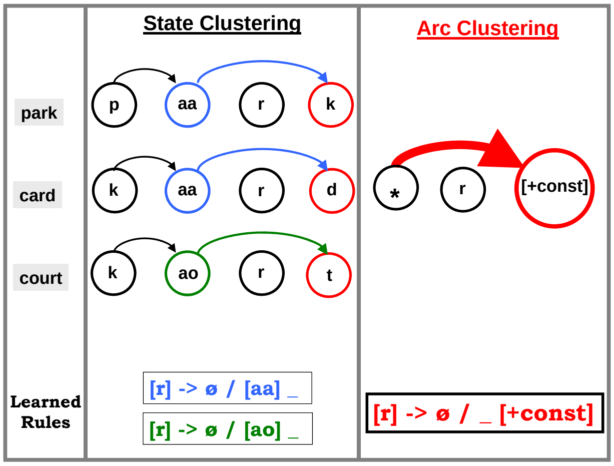


Fig. 2. *An example showing why arc clustering generalizes the R-dropping rule in British English better than not using arc clustering. In state-clustering (center column), two rules are learned: the first rule (in blue) specifies that when [aa] is followed by [r], [r] is deleted; the second rule (in green) specifies that when [ao] is followed by [r], [r] is deleted. Note that the crux of the rule – the right context of the deleted phone [r] – is not used to characterize the learned deletion rule. In contrast, a single rule, characterized by the right context of [r], is learned by arc clustering shown in red (right column)..*

Name	Tying Procedure	Section
Standard Tying	Triphone state clustering for *del, ins, typ*	II-C6
Refined Tying	(a) Arc clustering for *del* (b) Triphone state clustering for *ins* & *typ*	II-D2

TABLE I
TYING PROCEDURE FOR STANDARD TYING AND REFINED TYING.

Fig. 2 shows an example of how R-dropping [6] is modeled using triphone state clustering and arc clustering: After triphone state clustering, two rules are learned. The first rule (in blue) specifies that when [aa] is followed by [r], the [r] is deleted. The second rule (in green) specifies that when [ao] is followed by [r], the [r] is deleted. However, arc clustering results in one single rule represented (in red), specifying that [r] is deleted when followed by consonants. This rule learned from arc clustering is more general than the fragmented rules learned through state clustering.

2) Derivation for Refined Tying: Consider a triphone state $(s_{k-1} - s_k + s_{k+1})$. *Refined Tying* is performed by the following two steps: (a) We use arc clustering to determine which deletion arcs to tie together and then estimate the tied deletion probabilities accordingly; and (b) we estimate the

[8] The colon symbol ":" indicates that the vowel is elongated.

[9] [p-aa+r] represents a triphone where "−" denotes that monophone [aa] is preceded by [p] and "+" denotes that [aa] is followed by [r].

typical and insertion transition probabilities originating from s_k just as in the state-tying case, but with a new Lagrangian constraint, as the total sum of deletion probabilities leaving s_k are predetermined by arc clustering. The comparison between Standard Tying and Refined Tying is summarized in Table I.

(a) Arc Clustering for Deletion Transitions: Assume we are given an estimated HMM model $\lambda = \{A, B\}$, which we use to generate the most likely alignment between the given states and observations in the training data. Assume H_f is some feature that specifies the phonetic context of normal state x. We want to compute the likelihood of the normal state x given $x \in H_f$.

Given a normal state x, it can either be skipped or not skipped. The expected counts of state x being skipped given $x \in H_f$ is

$$C_{x_{skipped}} = \text{total soft counts of } x \text{ skipped when } x \in H_f.$$

The expected counts of state x not being skipped given $x \in H_f$ is

$$C_{x_{not-skipped}} = \text{total soft counts of } x \text{ not skipped when } x \in H_f$$

The likelihood of the normal state x, given $x \in H_f$ is

$$L(x|x \in H_f) = \left(\frac{C_{x_{skipped}}}{C_{x_{skipped}} + C_{x_{not-skipped}}}\right)^{C_{x_{skipped}}} \left(\frac{C_{x_{not-skipped}}}{C_{x_{skipped}} + C_{x_{not-skipped}}}\right)^{C_{x_{not-skipped}}}.$$

Similarly, $L(x|x \notin H_f)$ and $L(x)$ can also be obtained. The likelihood increase of the split is

$$\Delta \log L = \log \frac{L(x|x \in H_f) L(x|x \notin H_f)}{L(x)}. \tag{4}$$

Suppose that after arc clustering we obtain J groups of tied deletion arcs. Group j is specified by $D_j = (\sigma_j, \varsigma_j, \tau_j)$, where σ_j and τ_j specify the origin and target of the transition arc and ς_j specifies the skipped state. For example, when considering the deletion of [r], a possible clustered deletion transition arc might be σ specifying the phone before the deleted phone is a back vowel (e.g., [aa]), ς specifying the deleted phone is a rhotic liquid (i.e., [r]), and τ specifying the phone after the deleted phone is either a consonant (e.g., [t]) or silence.

All deletion transitions that originate from state $q_t \in \sigma_j$, skip state $d \in \varsigma$, and arrive at $q_{t+1} \in \tau_j$ are tied; this deletion probability is represented by the parameter $\overline{A_{D_j}}$:

$$\begin{aligned} \overline{A_{D_j}} &= \overline{P}(q_{t+1} \in \tau_j, r = del, d \in \varsigma_j | q_t \in \sigma_j) \\ &= \frac{\sum_{t=0}^{T} P(\mathbf{O}, q_t \in \sigma_j, r = del, d \in \varsigma_j, q_{t+1} \in \tau_j | \lambda, C)}{\sum_{t=0}^{T} \sum_r P(\mathbf{O}, r, q_t \in \sigma_j | \lambda, C)}, \end{aligned} \tag{5}$$

where d is the deleted (skipped) normal state. Clustered deletion probability A_{D_j} can be estimated using the Baum-Welch algorithm similarly to traditional HMM systems.

(b) Triphone State Clustering for Insertion and Typical Transitions: We previously estimated the deletion probability from arc clustering. Now we estimate the insertion and typical transition probability. After state clustering, we assume triphone states are clustered into I groups. Group i is specified by $G_i = (\zeta_\ell^i, \zeta_m^i, \zeta_r^i)$, where ζ_ℓ^i, ζ_m^i, and ζ_r^i specify the left-context, center, and right-context states. Similar to using Baum's auxiliary function in typical HMM systems but applying a new Lagrangian constraint, it can be shown that the tied typical and insertion transition probabilities are redistributed proportionally as

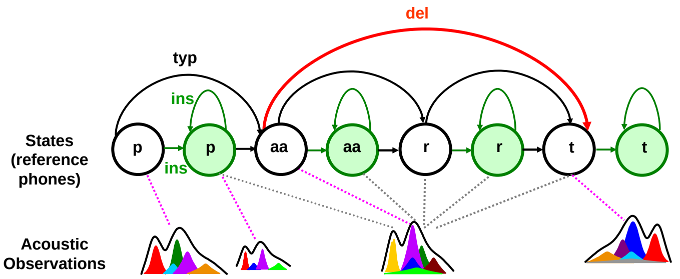


Fig. 3. *HMM network of acoustic pronunciation model (APM), the acoustic counterpart of PPM. The continuous acoustic observations are modeled by Gaussian mixture models. All other elements use the same symbol as in Fig. 1.*

$$\begin{aligned} \overline{A_{G_i,r}} &= \overline{P}(q_t \in G_i, r | q_t \in \zeta_m^i) \\ &= \frac{\sum_{t=0}^{T} P(\mathbf{O}, r, q_t \in G_i | \lambda, C)}{\sum_{t=0}^{T} \sum_{r \in R} P(\mathbf{O}, r, q_t \in \zeta_m^i | \lambda, C)} (1 - P_D(s_k)), \end{aligned} \tag{6}$$

where $R = \{typ, ins\}$ and $P_D(s_k)$ is the sum of all clustered deletion probabilities leaving the triphone state $(s_{k-1} - s_k + s_{k+1})$. Refer to Appendix A for details of how to obtain $P_D(s_k)$.

Table I summarizes the two tying structures mentioned: Standard Tying refers to the state clustering approach in Section II-C6 and Refined Tying refers to the approach introduced in Section II-D2.

E. Acoustic Pronunciation Model (APM)

1) Motivation: Some dialect differences are too subtle to elicit phonetic transformations. For example, in American English, the phone [p] is less aspirated when it is preceded by a obstruent consonant (e.g.,[s]). Therefore, the [p] in *pray* and *spray* acoustically differ in the amount of aspiration; the former aspirates more air than the latter. This is not necessarily true in other English dialects (e.g., Indian English). Since these dialect differences are at the sub-phonetic level, it is crucial to also characterize acoustic differences.

2) Model: To better characterize such subtle acoustic differences, we propose APM, the acoustic counterpart of PPM. APM can be derived from PPM by replacing the discrete observation probability $B_x(k)$ in Eq. (2) with a continuous *pdf* $B_x(\mathbf{z})$, modeled as a mixture of Gaussians:

$$B_x(\mathbf{z}) = P(O_t|q_t = x) = \sum_{l=1}^{M} w_{xl}\mathcal{N}(\mathbf{z}; \mu_{xl}, \mathbf{\Sigma}_{xl}), \qquad (7)$$

where $1 \leq t \leq T$ and T is the total number of time frames in the observation utterance; $\mathcal{N}$ is the normal density; $w_{xl}, \mu_{xl}, \mathbf{\Sigma}_{xl}$ are the mixture weight, mean vector, and covariance matrix of state x; and mixture l, $1 \leq x \leq N$, $1 \leq l \leq M$, $\sum_{l=1}^{M} w_{xl} = 1$. The HMM network of APM is shown in Fig. 3. The same tying as before (i.e., Standard Tying and Refined Tying) can be used, but in this work, we only show results using Standard Tying.

F. Remarks: Comparison with Other Models

Fig. 4 is a diagram showing how our proposed systems relate with other dialect recognition systems. GMM is often the basis of acoustic systems. Acoustic phonetic models (APM0) [13] stem out from GMM by modeling acoustic information in monophone categories. APM0 can be further extended to our proposed APM system using context clustering to refine acoustic characterization.

Though APM is akin to traditional GMM approaches, its underlying framework is designed to make rule interpretation and analysis intuitive to humans. On the other hand, models such as Phone-GMM-Supervector-Based SVM Kernel in [17] uses GMM as a basis to train discriminative classifiers, which focus more on dialect recognition than rule interpretation.

PRLM (Phone Recognition followed by Language Modeling) [9] represents a classic model in phonetic systems. BinTree language modeling [19] and our proposed PPM both refine PRLM by exploiting context clustering. In contrast to BinTree, PPM focuses on the interpretation of dialect-specific rules, instead of dialect recognition performance. In addition, their probability models are different: in BinTree language modeling, the probability of a current observation is conditioned on a cluster of past observations [19], whereas in PPM, the probability of a current observation is conditioned on a reference phone and its context. Using reference phones as a clean comparison basis, instead of noisy decoded phones as in PRLM and BinTree, we can explicitly characterize what kinds of dialect-specific transformations are occurring and how often they occur. This trait makes PPM stand out among others when it comes to interpreting dialect-specific rules.

Our proposed framework applies to both acoustic and phonetic approaches, making PPM and APM analogs of each other. APM can thus be interpreted in two different perspectives: (1) an extension of the monophone-based APM0 system and (2) the acoustic counterpart of PPM.

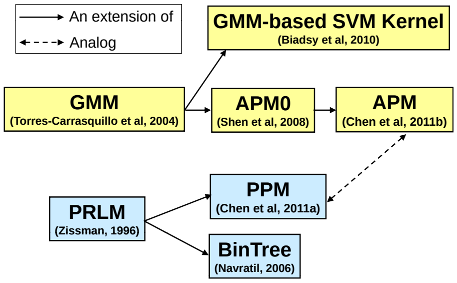


Fig. 4. *How PPM and APM relate to each other and other dialect recognition systems.*

III. Pronunciation Conversion Experiment

The objective of this experiment is to learn the phonetic rules that map American pronunciation (reference dialect) to British pronunciation (dialect of interest). To do so, we generate estimated British surface phones from American reference phones and evaluate how well they match the ground-truth British surface phones. It is assumed that the system that most accurately converts American pronunciation to British pronunciation learned the phonetic rules the best. This pronunciation conversion experiment shares a similar spirit to [28] in automatically generating dictionary pronunciations.

A. Experiment Design

1) Corpus: WSJ-CAM0 (Wall Street Journal recorded at the University of CAMbridge phase 0) [29] is the British English version of the American English WSJ0 corpus [30]. WSJ-CAM0 is read speech recorded in quiet background and sampled at 16 kHz. We used the data partition of WSJ-CAM0: training set is 15.3 hr (92 speakers); the test and development set are each 4 hr (48 speakers).

2) Phonetic Representation: The WSJ-CAM0 dataset includes ground-truth phonetic transcriptions using the extended Arpabet [31], which was designed to include specific symbols for British English. British and American pronunciation can thus be represented with the same set of phone symbols from the extended Arpabet [31]. We denote these ground-truth phonetic transcriptions as ground-truth British pronunciation surface phones O^*.

The reference phones C represent how a typical general American English speaker would have produced the utterance read by the British speakers. We therefore obtained reference phones C (also represented using the extended Arpabet) by converting the word transcriptions of WSJCAM0 to the most common pronunciation in the CALLHOME American English Lexicon (Pronlex) [32] without using any acoustic data[10]; i.e., the entry with the highest occurrence frequency in the dictionary was chosen. This frequency information is from forced alignments of the Fisher corpus [47] used in [34].

Manual inspection confirms that major pronunciation differences between American and British English described in [6] were represented in the dataset. For example, vowels in *trap* and *bath* are represented as [ae] and [aa] in the ground-truth surface phones O^*.

3) Pronunciation Conversion System Setup: The reference phones C and ground-truth surface phones O^* in the training set were used to train different variants of the PPM systems (monophone, standard tying, refined tying) described in Section II.

Fig. 5 shows the system diagram during test time: given the reference phones C in the test set and a trained PPM

[10] Since WSJCAM0 are recordings from British speakers, it is inappropriate to use the acoustic data to derive the American pronunciation.

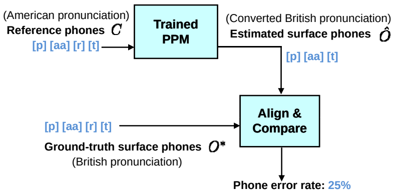


Fig. 5. *Procedure of pronunciation conversion experiment.*

system, the PPM system generates the most likely observations $\hat{O}$ (converted British pronunciation). Dynamic programming is used to align estimated surface phones $\hat{O}$ with the ground-truth surface phones O^* to compute phone error rate. We also computed a baseline where no pronunciation model was used: the ground-truth surface phones O^* (British pronunciation) were directly compared against the reference phones C (American pronunciation).

4) Statistical Test: We used the matched pairs test [35] to assess if the performance differences between two systems were statistically significant. Each type of error (deletion, insertion, and substitution) was analyzed separately.

5) Remarks: In this experiment, we expect to learn the R-dropping rule, which is characteristic of certain dialect groups in British English (e.g., received pronunciation (RP) [6]). When considering more than just the RP dialect, the R-dropping rule can still be learned if the rule is still more prevalent than in American English, which is the case in this experiment. The probability of the rule occurring would be lower than when considering merely the RP dialect. Note that even if all the British speakers in WSJCAM0 belong to the RP dialect, it is not necessarily true that the R-dropping rule will always be followed since these phonetic rules are probabilistic, not hard-and-fast binary rules. Note also that if there is an utterance in the test set that does not follow the R-dropping rule, and if our algorithm predicts a deletion of /r/, this is counted as an error in the experiment.

B. Results and Analysis

1) Phonetic Context Improves Performance: The phone error rates for the baseline and variants of PPM are summarized in Table II. The monophone PPM (System S_1) outperforms the baseline (System S_0) by 30.4% relative. Relative gains from triphone PPMs (System S_2 and S_3) are even greater, both reducing the baseline phone error rate by 58.5% relative. Both triphone PPM systems (System S_2 and S_3) outperform the monophone PPM system (System S_1) by 40.4% relative, indicating the importance of phonetic context in modeling dialect differences. All differences in performance are statistically significant ($p < 0.001$).

2) Refined Tying Models Deletions Better: Table II shows that the Standard Tying Triphone System S_2 shows an increase in deletion errors (5% relative, $p < 0.001$) compared to Monophone PPM (System S_1). This suggests that Standard Tying over-generalizes deletion rules. In contrast, the Refined Tying Triphone System S_3 improves the deletion errors of the monophone PPM system by 43% relative. When considering deletion errors, Refined Tying outperforms Standard Tying by 33% relative. These results support our intuition that arc clustering is suitable in modeling deletions and is consistent with the linguistic knowledge that a phone is generally affected more by its right-context than left-context; e.g., R-dropping in British English [6]. Among the [r]'s that were incorrectly deleted (i.e., [r] was deleted in the converted British pronunciation $\hat{O}$ but retained in the ground-truth British pronunciation O^*) in Standard Tying, Refined Tying correctly generated 24% of these [r]'s.

Although Refined Tying (System S_3) reduces deletion errors, its insertion errors increase when compared to Standard Tying (System S_2). This phenomenon might be caused by data sparsity, since Refined Tying requires more model parameters. The matched pairs test shows that these two systems make statistically different errors ($p < 0.001$), implying that these two systems could complement each other in learning rules.

C. Summary

In this section, we assess the rule learning ability of PPM by examining how well it converts pronunciation from one dialect to another. We show that phonetic context is important for conditioning rules and Refined Tying is more suitable for characterizing deletion rules than Standard Tying. In this experiment, we analyzed dialect differences in read speech between British and American English. In the next set of experiments, we examine dialect differences in conversational speech in American English.

IV. Dialect Recognition Experiment

In this section, we use dialect recognition as a tool to help us analyze how well the proposed models explicitly characterize rules. We want to understand whether the proposed models can still achieve similar performance to baseline systems though they are designed to explicitly characterize rules instead of optimize dialect recognition performance. In addition, since the proposed models characterize dialect-specific information differently from standard dialect recognition techniques, we want to assess if combining system scores can lead to fusion gains.

A. Corpus

To the best of our knowledge, there is virtually no corpus suitable for automatic dialect analysis. We originally intended to use NIST LRE data [15], but realized the NIST data is not suitable for developing rule-learning algorithms. Word transcriptions and pronunciation dictionaries specifying dialect variation are essential to establishing a well-defined reference pronunciation for explicitly analyzing dialect differences. Unfortunately, the majority of the NIST data do not possess these properties. In addition to the NIST LRE data, we have also investigated over a dozen dialect corpora [36], but virtually none were suitable for large-scale automatic dialect analysis.

Consequently, we constructed a corpus from StoryCorps (raw data from [37]), consisting of (a) African American Vernacular English (AAVE) spoken by self-reported African Americans and (b) non African American Vernacular English (Non-AAVE) spoken by self-reported white Americans. The dataset was designed with the following desirable qualities:

System	Total Error (%)	Del. Error (%)	Ins. Error (%)	Sub. Error (%)
S_0 Baseline	21.7	4.0	3.6	14.2
S_1 Monophone PPM	15.1	2.0	3.3	9.8
S_2 Standard Tying Triphone PPM	9.0	2.1	1.9	5.0
S_3 Refined Tying Triphone PPM	9.0	1.4	2.6	5.0

TABLE II

PPM system performance (in phone error rate) in converting American pronunciation to British pronunciation. Del: deletion; Ins: insertion; Sub: substitution.

	AAVE	non-AAVE
Training	12.6 hr (43 spkrs)	9.8 hr (26 spkrs)
Development	4.2 hr (17 spkrs)	3.0 hr (11 spkrs)
Test	4.1 hr (14 spkrs)	2.8 hr (14 spkrs)

TABLE III

Duration and speaker number breakdown of the two dialects used in StoryCorps Corpus.

1) Conversational speech, since natural and spontaneous speaking style elicits non-mainstream dialects more easily [6].
2) Only conversations between friends or family members of the same dialect were chosen to minimize accommodation issues. For example, this issue arises when an AAVE speaker (subconsciously) suppresses AAVE dialect characteristics when speaking with non-AAVE speakers [38], [39].
3) Gender [6] and age [6] were controlled across all data partitions.
4) Consistent channel conditions across recording sites and high-quality recording (e.g., 16 kHz sampling rate).
5) Word-transcribed and sufficiently large enough to train dialect recognition models with standard statistical techniques [36].
6) The dialect of interest, African American Vernacular English (AAVE), is a relatively well-studied dialect in American English (e.g., [45], [46]).
7) Preliminary listening tests showed that native American English speakers can perceive dialect differences.

The conversations are about the individuals' life stories. The topics cover a wide variety, including love stories of how couples met, grandparents' experiences in war, and struggles people faced climbing up the social and economic ladder.

B. Training Phase

We consider the following baseline systems: GMM [10], acoustic phonetic model (APM0) [13], and Parallel Phone Recognition followed by Language Modeling (PPRLM) using adapted tokenizers [13], referred to as PRLM in the rest of the paper. In each of these systems, the goal of training is to estimate separate models of AAVE and non-AAVE:

1) GMM: Each GMM has 2048 mixture components. Experimental setup is similar to [10]. The front-end features are shifted delta cepstrum (SDC). A universal background model was first trained on the training data of all dialects (AAVE and non-AAVE), and then the means of the dialect-specific GMMs were separately adapted to the AAVE and non-AAVE data using Maximum A Posteriori (MAP) [40].

2) APM0: We denote adapted phonetic models [13] as APM0 to avoid confusion with our proposed APM (acoustic pronunciation model). We can think of APM0 as a simplified version of APM with two differences: (1) APM0 is monophone-based and (2) APM0 uses a root phone recognizer (rather than forced alignment with word transcriptions) to tokenize the speech signal to phonetic units (phone loop decoding). We first train a general-purpose root phone recognizer on the WSJ0 corpus using the same acoustic features (perceptual linear predictive (PLP) coefficients [41], 1st delta features, and 2nd delta features) as in [13]. We used this general-purpose WSJ0 root phone recognizer to decode the StoryCorps dataset into 40 monophone classes. A universal background model was trained for each of 40 monophone classes using the StoryCorps data and then adapted to dialect-specific data using 32 Gaussians/state, resulting in an APM0 adapted phone recognizer.

3) PRLM: In the PRLM (Parallel Phone Recognition followed by Language Modeling) system, we used the APM0 adapted phone recognizer (from Section IV-B2) to tokenize phone-lattices to train separate AAVE and non-AAVE trigram language models, as in [13]. The phone-lattice encodes uncertainties in the phone decoding by including not just the 1-best phone sequences, but also the lower ranking sequences.

4) PPM: Fig. 6 shows the training procedure using surface and reference phones as features. Reference phones were obtained through forced alignment using the word transcripts and the general-purpose WSJ0 root phone recognizer. We applied the general-purpose WSJ0 root phone recognizer (from Section IV-B2) to decode the 1-best surface phones from the StoryCorps data using phone-loop grammar. This is in contrast to the phone lattices used by PRLM. We used the extracted surface and reference phones to separately train AAVE and non-AAVE monophone PPM (MonoPPM) systems. The dialect-specific MonoPPM systems were used to initialize the training of dialect-specific triphone systems. Both Standard Tying (PPM1) and Refined Tying (PPM2) listed in Table I were separately applied to the trained dialect-specific triphone systems, resulting in Tri-PPM1 and Tri-PPM2 respectively.[11]

5) APM: The training for APM is similar to PPM, resulting in dialect-specific MonoAPM systems. The only difference is that acoustic observations (i.e., front-end features as in APM0), instead of surface phones, were used. Standard Tying was used to cluster triphone states, resulting in dialect-specific Tri-APM systems.

[11] The mainstream phonotactic approaches in language and dialect recognition do not adapt the phone recognizers. We follow this convention for PPM.

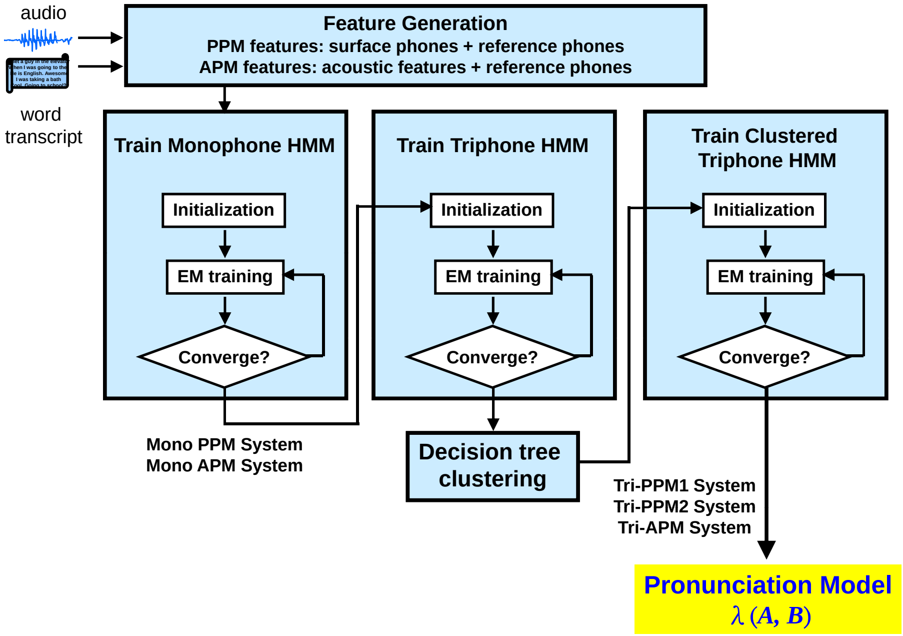


Fig. 6. *Training procedure for PPM and APM.*

The training setup is summarized in Table IV. We see that the different algorithms are not matched in multiple dimensions. For example, GMM and APM variants (APM0, MonoAPM and Tri-APM) use acoustic features, while PRLM and PPM variants (monoPPM, Tri-PPM1 and Tri-PPM2) use phonetic features. We also note that while PRLM uses surface lattice phones, all PPM variants use 1-best surface phones. Our rationale is that the baseline methods (GMM, APM0 and PRLM) are well-established methods in the field, and so we opted to match the setup used in the original publications for consistency, rather than modify these published algorithms to match every aspect of our proposed model's implementation.

We also include the monophone-based systems (APM0, MonoPPM, MonoAPM) to understand the trade-off between computational complexity and the rule prediction power.

C. Recognition Phase

The developmental and test sets were segmented into 30-sec trials. After training, each of the systems has separate models of AAVE and non-AAVE. We used the likelihood ratio test to perform dialect recognition on the trials.

1) Likelihood Ratio Test: Similar to language recognition, dialect recognition is viewed as a detection task. Dialect decisions are made via a likelihood ratio test. Given the test utterance O, the target dialect model λ_{AAVE}, the reference dialect model $\lambda_{non-AAVE}$ and a decision threshold θ, we infer that test utterance O was produced by the AAVE dialect if

$$\frac{1}{T}\log\frac{P(O|\lambda_{AAVE})}{P(O|\lambda_{non-AAVE})} > \theta, \quad (8)$$

where T is the normalization constant. In PPM, T refers to the number of surface phones in observation O. In all other systems, T denotes the duration of the test utterance.

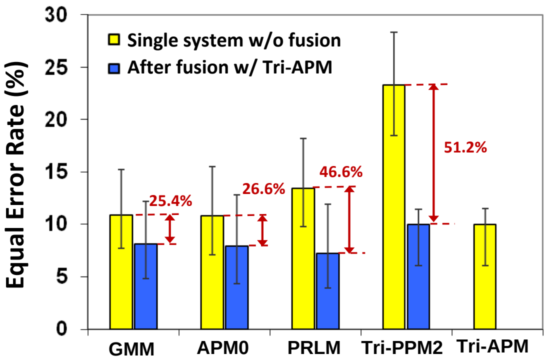


Fig. 7. *Dialect recognition performance before and after baseline systems fuse with APM. Light yellow bars: single systems; blue bars: fusion with proposed APM. The error bars are 95% confidence interval; confidence intervals are not necessarily symmetric.*

2) Fusion: Fusion is widely used in speaker and language recognition tasks to combine system outputs [42]. Here, we used a Gaussian classifier with a diagonal covariance matrix [43] to calibrate and fuse the score outputs from individual systems. The developmental set was used to tune the weights of the fusion classifier.

D. Results and Discussion

1) Individual Systems: Table IV summarizes the dialect recognition results. We observed the following:

- All acoustic systems (GMM, APM0, MonoAPM and Tri-APM) have similar performance and outperform phonetic systems (PRLM, monoPPM, Tri-PPM1 and Tri-PPM2).
- The best APM system (Tri-APM) outperforms the best PPM system (Tri-PPM2) by 57.2% relative. Among phonetic systems, PRLM is the best.
- Phonetic context improves performance: The triphone systems outperform their monophone counterparts for

	System	PRLM	Mono PPM	Tri PPM-1	Tri PPM-2	GMM	APM0	Mono APM	Tri APM
Training setup	Features	surface phone lattices	surface phones	surface phones	surface phones	SDC	PLP	PLP	PLP
	Modeling unit	trigram	monophone	triphone	triphone	N/A	monophone	monophone	triphone
	Word transcripts	N	Y	Y	Y	N	N	Y	Y
	WSJ0 root recognizer	Y	Y	Y	Y	N/A	Y	N/A	N/A
	Tying structure	N/A	N/A	standard	refined	N/A	N/A	N/A	standard
Dialect recognition	Equal error rate (%)	13.45	26.8	24.33	23.3	10.56	10.78	10.76	9.97

TABLE IV
COMPARISON OF TRAINING AND DIALECT RECOGNITION RESULTS

both PPM and APM. In particular, Tri-PPM1 and Tri-PPM2 outperform MonoPPM by 9.2% and 13.1% relative, respectively. Tri-APM outperforms MonoAPM by 8.1% relative.
- Refined Tying outperforms Standard Tying: Tri-PPM2 (Refined Tying) compares favorably with Tri-PPM1 (Standard Tying): 4.4% improvement relative. These results are consistent with the pronunciation conversion experiment in Section III.

2) Fusion with APM: We chose the best performing APM system (Tri-APM) to fuse with other systems. The results are shown in Fig. 7, suggesting that Tri-APM complements other systems. We discuss this in more detail below.

- There is at least 25% relative gain when Tri-APM is fused with GMM or APM0 (Fig. 7). The fused systems performed better than individual systems, suggesting that the error patterns of APM are different from other acoustic systems. This additional gain may have resulted from explicitly modeling dialect-specific rules. Note that while the fused results of GMM and APM0 are within the 95% confidence interval of their unfused counterparts, they are very near the lower boundary of the confidence interval. This indicates that the performance difference is close to statistical significance ($p \approx 0.05$).
- The fusion of Tri-APM with PRLM produced more than 46% relative gain. The fused results of PRLM are outside the 95% confidence intervals of their unfused counterparts, indicating the unfused and fused results are statistically significant ($p < 0.05$). This result is expected: since PRLM is a phonotactic system, it should complement Tri-APM more than other acoustic systems.
- We expected Tri-PPM2 to also fuse well with Tri-APM. However, while there is a relative gain of 57.2% when compared to Tri-PPM2's original performance ($p < 0.05$), the fused result is similar to that of Tri-APM alone. We observed that Tri-PPM2 scores correlate with those of Tri-APM. We think this is because the acoustic characterizations from Tri-APM already incorporates the phonotactic information modeled by Tri-PPM2, causing no fusion gains from Tri-APM's standpoint.

3) Why APM Performs Better Than PPM: Although the proposed topology (Fig. 1) models the three types of phonetic transformations explicitly, only using 1-best surface phones in PPM may limit its ability to fully model subtle acoustic characteristics of these transformations. This might explain why lattice-based PRLM performs better than PPM. The use of lattice surface phones in PPM is outside the scope of this paper. We leave this extension to future work.

4) Improving APM: Since APM is an ASR-inspired model, we used PLP features. SDC features have been reported to outperform traditional cepstral features in language and dialect recognition tasks [44], [10]. Incorporating SDC features in APM will be explored in the future.

E. Summary

In this section, we showed that despite the additional constraints of explicitly modeling dialect-specific rules, the proposed APM system (i.e., Tri-APM) obtains comparable dialect recognition performance to classic systems. These results suggest that dialect differences can be characterized as rules and used to recognize dialects. In addition, the fusion gains received from fusing APM with baseline systems imply that APM is exploiting dialect-specific information not modeled by conventional systems. This complementary information likely results from the explicit characterization of dialect-specific rules. In the next section, we examine what these rules are, how well the different systems are able to retrieve these rules, and how the proposed system can refine linguistic rules.

V. REGION-OF-INTEREST RETRIEVAL EXPERIMENT

In this section, we evaluate how well dialect recognition systems retrieve regions of dialect differences defined by linguists. We use the same StoryCorps data described in Section IV-A.

A. Setup

1) Linguistic Rules: A set of phonetic rules describing the AAVE dialect was developed by a team of collaborating linguists.[12] The development involved consulting the literature (e.g., [45], [46]), consulting with other phoneticians, and acoustic/perceptual analysis on a small set of transcribed data from the Mixer corpus [47]. Rules that did not occur in the StoryCorps data were excluded; the rest are listed in the 2nd column of Table V. Note that these phonetic rules were the best approximations the linguists came up with using the Arapbet phone set. For example, in row 4 of Table V, the degliding of [ay] might result in a phone close to [aa], a phone close to [ae], or a phone with acoustic characteristics similar to a phone in

[12] Consultation of linguistic experts was necessary to bridge the gap between existing literature and practical needs: (1) As dialects evolve with time, some linguistic literature might be out-of-date. (2) Linguistic literature might not always have comprehensive documentation of what is important for engineering applications (e.g., exhaustive list of rules with frequency occurrence statistics).

Phone of Interest	Linguistic rules for AAVE Dialect	Example
[r]	[r] $\rightarrow \emptyset$/ [+vowel] __ [+vowel] [r] $\rightarrow \emptyset$/ [+vowel] __ [r] $\rightarrow \emptyset$/ [+consonant] __ [ao/ow/oy]	or anything more throw
[ng]	[ng] $\rightarrow$ [n]	thing
[ay]	[ay] $\rightarrow$ [ae $\Vert$ aa] / __ [+consonant]	like
[eh]	[eh] $\rightarrow$ [ih] / __ [+nasal] [eh] $\rightarrow$ [ey] / __ [+liquid]	then air
[l]	[l] $\rightarrow \emptyset$/ [+vowel] __[+consonant] [l] $\rightarrow \emptyset$/ [uw$\Vert$uh] __	all the time cool
[dh]	[dh] $\rightarrow$ [v$\Vert$d] / [+vowel] __ [+vowel]	brother
[ih]	[ih] $\rightarrow$ [ae $\Vert$ eh] / [th] __ [ng]	thing
[aw]	[aw] $\rightarrow$ [ow $\Vert$ ay $\Vert$ aa $\Vert$ uw] / __ [l] [aw] $\rightarrow$ [ow $\Vert$ ay $\Vert$ aa $\Vert$ uw] / __ [t]	owl about

TABLE V
LINGUISTIC RULES FOR THE AAVE DIALECT PROPOSED BY PHONETICIANS. || REPRESENTS LOGICAL "OR"; THE COMMA SYMBOL "," REPRESENTS LOGICAL "AND". THE PROBABILITY OF THE RULE OCCURRING GIVEN THE PHONE OF INTEREST IS SHOWN IN FIG. 8.

between [ae] and [aa], thus both transformations are included. In this section, we treat these linguistic rules as ground-truth that we attempt to recover.

2) AAVE-extended American Pronunciation Dictionary: We use the linguistic rules to augment the WSJ0 American English pronunciation dictionary (from Section III-A2) to include AAVE pronunciations, which we denote as the WSJ0 AAVE-extended American pronunciation dictionary.

3) Target vs. Non-Target Trials: Reference triphones were obtained through forced alignment using word transcripts and the general-purpose WSJ0 root phone recognizer (like the training of PPM in Section IV-B4). Ground-truth surface phones were obtained through forced alignment using the word transcripts, the general-purpose WSJ0 root phone recognizer and the AAVE-extended American pronunciation dictionary. The reference and surface phones were aligned using dynamic programming. The speech segment corresponding to each reference triphone is considered a trial. A *target trial* occurs when a reference triphone and corresponding surface phone match the linguistic rules. All other trials are *non-target trials*. Examples are shown in Table VI.

While we plan to treat the linguistic rules from Table V as ground-truth rules we attempt to recover, we recognize that these rules are not hard-and-fast rules that always occur and that there might be potential oversights from manual linguistic analyses. For example, while row 3 in Table V predicts the degliding of [ay] in AAVE, this might only occur sometimes in real AAVE speech. To quantify the validity of these linguistic rules with our dataset, Fig. 8 shows the probability of the phonetic transformations occurring in the specified phonetic context (specified in Table V) given the phone of interest in AAVE and non-AAVE data, respectively. We observe that while these rules were devised for AAVE speech, they also occur in non-AAVE speech. We also observe that the rules occur more frequently (except for [ay]) in AAVE data than in non-AAVE data, suggesting that the linguistic rules in Table V are generally informative of AAVE and non-AAVE differences. The exception of [ay] suggests that the [ay] linguistic rule can be potentially refined to better manifest dialect characteristics. We will come back to this issue in Section V-C.

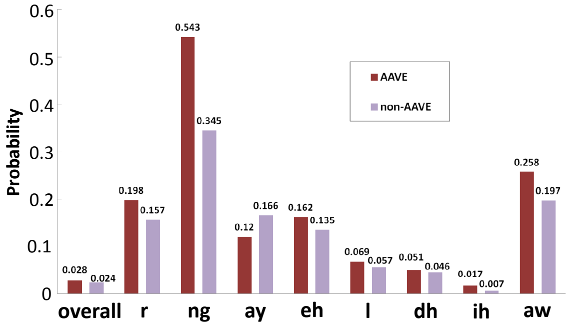


Fig. 8. *Probability of the AAVE linguistic rule occurring given the phone of interest in AAVE and non-AAVE speech.*

4) Retrieving Target Trials: For each AAVE trial (i.e., speech segment corresponding to a reference triphone), we perform a log likelihood ratio (LLR) test. Let O be the (acoustic or phonetic) observation of the trial obtained like in Section IV. Let T be the duration of O. Therefore, in PPM, T is the number of observed surface phones. In all other systems, T is the duration of the reference triphone. Given a decision threshold θ, we infer that the trial is a target trial if

$$\frac{1}{T}\log\frac{P(O|\lambda_{AAVE})}{P(O|\lambda_{non-AAVE})} > \theta, \tag{9}$$

where λ_{AAVE} and $\lambda_{non-AAVE}$ are the trained models of all dialect recognition systems in Table IV. The decision threshold was tuned on the developmental set.

5) Standard Retrieval Metric: Recall is the proportion of ground-truth target trials that are successfully retrieved. *Precision* is the proportion of retrieved trials that are ground-truth target rules. In these experiments, we use the F-measure to represent the overall retrieval performance of a system. The *F-measure* is the harmonic mean of recall and precision: $F = \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$.

6) Baseline F-measure: We set the precision level to the proportion of target trials in the test set, which represents the chance level precision rate. Given this precision rate, we obtain the value of recall that leads to the optimal F-measure. This F-measure denotes the baseline.

B. Results

The retrieval results are summarized in Table VII. The F-measure at chance level is 0.0547, which is similar to the results of PRLM, GMM, APM0. Even though Tri-APM performs similar to other acoustic systems in dialect recognition tasks, its retrieval performance outperforms others (6.69%, 7.26% absolute gains when compared to GMM and APM0). Similarly, although Tri-PPM2 performs worse than PRLM in dialect recognition, its retrieval result compares favorably (2.1% absolute gain) to PRLM. These results suggest that explicit modeling of rules helps locate dialect-specific regions.

C. Discussion: Learned Rules Refine Linguistic Rules

In Section V-A3, for analysis purposes we assumed that the AAVE linguistics rules are the ground-truth when conducting

Reference Triphone	Ground-truth Surface Phone	Target/Non-Target Trial?
[eh+n]	[ih]	Target Trial because of phonetic rule [eh] —>[ih] ⁄ __ [+nasal]
[eh+g]	[ih]	Non-Target Trial because it does not match any linguistic rules
[eh+n]	[eh]	Non-Target Trial because it does not match any linguistic rules

TABLE VI
EXAMPLES OF TARGET AND NON-TARGET TRIALS.

System	PRLM	MonoPPM	Tri-PPM1	Tri-PPM2	GMM	APM0	MonoAPM	Tri-APM
F-measure	0.0515	0.0640	0.0697	0.0725	0.0593	0.0536	0.0876	0.1262

TABLE VII
REGION-OF-INTEREST RETRIEVAL COMPARISON. F-MEASURE BASELINE (AT CHANCE LEVEL) IS 0.0547.

	Linguistic rules for AAVE Dialect	Example	Top ranking triphones	Example
[r]	$[r] \rightarrow \emptyset /$ [+vowel] __ [+vowel] $[r] \rightarrow \emptyset /$ [+vowel] __ $[r] \rightarrow \emptyset /$ [+consonant] __ [ao/ow/oy]	or anything more throw	[+diphthong] - [r] + [+vowel]	tiring
[ng]	[ng] → [n]	thing	[ih]-[ng]+[-silence]	seemingly
[ay]	[ay] → [ae || aa] / __ [+consonant]	like	[-glide] - [ay] + [+nasal] [ay] + [w] [-liquid, -central] - [ay] + [+glide]	my mom I was that I had
[eh]	[eh] → [ih] / __ [+nasal] [eh] → [ey] / __ [+liquid]	then air	[-voiced, -vowel] - [eh] + [n] [+voiced] - [eh] + [n]	pen when
[l]	$[l] \rightarrow \emptyset /$ [+vowel] __[+consonant] $[l] \rightarrow \emptyset /$ [uw||uh] __	all the time cool	[-back, -[ah]] - [l] + [+voiced, -sonorant, -vowel]	all but
[dh]	[dh] → [v||d] / [+vowel] __ [+vowel]	brother	[+vowel] - [dh] +[er]	brother
[ih]	[ih] → [ae || eh] / [th] __ [ng]	thing	[th] - [ih] + [ng]	thing
[aw] [aw]	[aw] → [ow || ay || aa || uw] / __ [l] [aw] → [ow || ay || aa || uw] / __ [t]	owl about	[aw] + [-nasal, +consonant] [-nasal, +consonant]	howl out

TABLE VIII
COMPARISON BETWEEN LINGUISTIC AND LEARNED RULES. LEFT: LINGUISTIC RULES FOR THE AAVE DIALECT PROPOSED BY PHONETICIANS. MIDDLE, RIGHT: TOP-RANKING TRIPHONES FROM APM USING LOG LIKELIHOOD RATIO (LLR). THE TOP-RANKING TRIPHONES OFTEN HAVE MORE REFINED PHONETIC CONTEXT THAN THE LINGUISTIC RULES. "||" REPRESENTS LOGICAL OR; THE COMMA SYMBOL "," REPRESENTS LOGICAL AND.

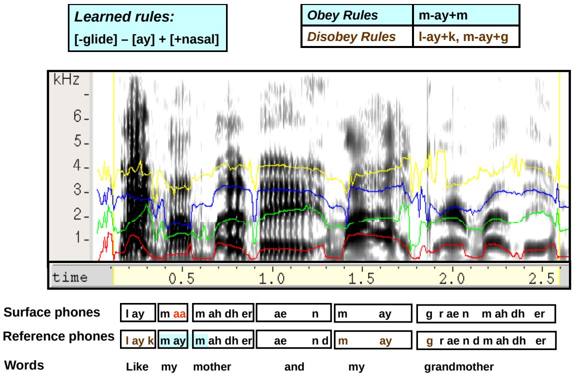


Fig. 9. *Learned rules help refine existing rules: [ay] does not de-glide when followed by any consonant; [ay] de-glides when followed by a nasal consonant.*

the retrieval experiments, though we observed that the linguistic rules might not be perfect (e.g., the rule for [ay] does not predict the AAVE dialect in Fig. 8). In this section, we explore and discuss the possibility of refining linguistic rules by comparing them with automatically learned rules.

In Tri-APM, dialect information is characterized as triphone models. For example, assume an acoustic observation O corresponds to a reference triphone [th-ih+ng]. If likelihood score of O given the AAVE [th-ih+ng] model, $P(O|\lambda^{\text{th-ih+ng}}_{\text{AAVE}})$, is high, but the likelihood score of O given the non-AAVE [th-ih+ng] model, $P(O|\lambda^{\text{th-ih+ng}}_{\text{non-AAVE}})$, is low, this triphone [th-ih+ng] specifies where dialect-discriminating information is rich. This triphone model acts as a rule learned from Tri-APM, because it is able to predict which phonetic conditions AAVE and non-AAVE speech possess different acoustic characteristics.

Therefore, to assess the most dialect-discriminating learned rules, we rank triphones according to their log likelihood ratio (LLR): For each occurring triphone in the test set of AAVE trials, its duration-normalized LLR between AAVE and non-AAVE models was computed. The average LLRs for each clustered triphone group were then ranked.

Table VIII shows the top ranking triphones (4th column) in descending order and compares them with the linguistic rules. We see that the top-ranking triphones often have more refined phonetic context than linguistic rules. For example, while the linguistic rule says that [ay] de-glides when preceding consonants, our system suggests that [ay] only de-glides when preceding certain consonants (e.g., [m].) Fig. 9 shows such an example, where there are three instances of [ay] followed by a consonant, but only one of them is de-glided. Therefore, phoneticians can potentially use learned rules to further investigate if existing rules need to be refined.

VI. DISCUSSION: MODEL ASSUMPTIONS, POSSIBLE EXTENSIONS, AND POTENTIAL APPLICATIONS

In this work, we focus on automatically learning pronunciation differences across English dialects. The proposed model assumes that the writing system between the different dialects are easily convertible and word transcriptions are available during training. Given a word, the model assumes speakers from different dialect groups might pronounce the word differently either at the acoustic or phonetic level. It is assumed that these differences can be probabilistically characterized as pronunciation patterns, which are denoted as phonetic rules.

In this work, we only condition these rules on the immediate left and right neighboring phones of the phone of interest. This conditioning can be easily extended so that phonetic rules can be potentially modeled more elegantly. Besides phonetic context, other information, such as syllable-level position and word boundaries, could also be further incorporated in modeling phonetic rules.

While we only show results on English dialects, the proposed model has also had some success on Arabic dialects [16].[13] In addition, our model can potentially be applied to other languages such as Spanish dialects (spoken in different regions of Spain and Latin America) and Mandarin dialects (spoken in China, Singapore and Taiwan). For example, phonetic rules characteristic of Caribbean Spanish include syllable-final /s, f/ debuccalized to [h], word-final /n/ velarized [48]. In Taiwanese Mandarin, schwa becomes backed when followed by a velar nasal and preceded by a labial consonant: [eng] → [ong]/____[b,p,f,m] [49], [50], where the phonetic symbols are in hanyu pinyin [51]. We have in fact obtained encouraging results on preliminary experiments on these languages, though we were hampered by limitation in available resources to continue our efforts.

Since our model only focuses on acoustics and phonetics, it is not suitable for analyzing dialect differences caused by higher linguistic levels such as syntax and prosody. However, our framework is not limited to analyzing dialects. It can be applied to characterizing pronunciation variation across different types of speaker groups, be they non-native accents, speech disorders (articulatory or phonological disorders), or individual speakers (e.g., speaker recognition). For non-native accents, ground-truth phonetic rules could be hypothesized by comparing the phonological systems in the first and second language, but determining how non-native a speaker is might require non-trivial perceptual evaluation. For speech disorders and speaker recognition applications, it might be useful to extend the framework to characterize pronunciation variation using limited data. These issues remain open research questions worthy of future investigation.

VII. Conclusion

We proposed a framework that refines the hidden Markov model to explicitly characterize dialect differences as phonetic and acoustic transformations, which are interpreted as dialect-specific rules. We demonstrated that the proposed PPM system is able to convert American English pronunciation to British pronunciation using learned rules. We also showed that the proposed systems can recognize different varieties of American English. In addition, the proposed APM system is able to retrieve and refine linguistic rules. The proposed framework of automatically characterizing pronunciation variation explicitly is potentially useful in fields such as sociolinguistics and speech pathology, which currently rely heavily on manual analysis.

Appendix A
Derivation Details of Refined Tying

The clustered probability for insertion and typical transitions in Eq.(6) in Section II-D2 depends on $P_D(s_k)$:

$$\begin{aligned}
P_D(s_k) &= \sum_j P(r = del | s_{k-1}, s_k \in \sigma_j, s_{k+1} \in \varsigma_j) \\
&\equiv \sum_j P(r = del | s_k \in \sigma_j, s_{k+1} \in \varsigma_j) \qquad (10) \\
&= \sum_j P(s_{k+2} \in \tau_j, r = del | s_k \in \sigma_j, s_{k+1} \in \varsigma_j) \\
&= \sum_j \frac{P(s_{k+2} \in \tau_j, r = del, d = s_{k+1} \in \varsigma_j | s_k \in \sigma_j)}{P(s_{k+1} \in \varsigma_j | s_k \in \sigma_j)} \\
&= \sum_j \frac{A_{D_j}}{P(s_{k+1} \in \varsigma_j | s_k \in \sigma_j)}, \qquad (11)
\end{aligned}$$

where $P(s_{k+1} \in \varsigma_j | s_k \in \sigma_j)$ is a bigram probability that can be obtained empirically from the training set. The $\equiv$ symbol in Eq. (10) is used due to our modeling assumption that the deletion of s_{k+1} only depends on the phone before and after (i.e, s_k ands_{k+1}).

Acknowledgments

The authors would like to thank the team of collaborating linguists who helped derive AAVE-specific phonetic rules: Chris Cieri (Linguistic Data Consortium, University of Pennsylvania), Robin Dodsworth (North Carolina State University), Arvilla Payne-Jackson (Howard University), Reva Schwartz (United States Secret Service), Stephanie Strassel (Linguistic Data Consortium, University of Pennsylvania), Richard Wright (Howard Uinveristy), Julie Vonwiller (Appen Bulter Hill Corporation), and Malcah Yaeger-Dror (University of Arizona). In addition, the authors would like to thank Adam Albright, Jim Glass, Pedro Torres-Carrasquillo, Tom Quatieri, and the anonymous reviewers for their insightful feedback.

References

[1] M. Wakelin, *Discovering English Dialects*, Vol. 235. Shire Publications Ltd., Oxford, UK, 2008.

[2] P. Ladefoged *Vowels and Consonants*, Blackwell Publishing, 2005.

[3] W. Labov, S. Ash, and C. Boberg, *The Atlas of North American English: Phonetics, Phonology, and Sound Change*, Mouton de Gruyter, Berlin, 2006.

[4] J. A. Goldsmith "Phonological Theory," in J. A. Goldsmith, *The Handbook of Phonological Theory*, Blackwell Handbooks in Linguistics, Blackwell Publishers, 1995.

[5] B. Hayes, *Introductory Phonology*, Blackwell Textbooks in Linguistics, Wiley-Blackwell, 2009.

[6] J. Wells, *Accents of English: Beyond the British Isles*, Cambridge University Press, 1982.

[7] K. Evanini, S. Isard, and M. Liberman, "Automatic formant extraction for sociolinguistic analysis of large corpora," Proc. Interspeech, 2009.

[8] P. Rose, *Forensic Speaker Identification*, Taylor and Francis, 2002.

[9] M. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, 4(1):31-44, 1996.

[10] P. Torres-Carrasquillo, T. Gleason, and D. Reynolds, "Dialect Identification using Gaussian Mixture Models," Odyssey: The Speaker and Language Recognition Workshop, 2004.

[13] Our later experiments reveal that in addition to phonetic rule differences, Arabic dialects also differ greatly at the vocabulary level. Therefore, dialect recognition performance can be even greater if all sources of differences are modeled.

[11] P. Angkititrakul and J. Hansen, “Advances in phone-based modeling for automatic accent classification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 634-646, 2006.

[12] R. Huang and J. Hansen, “Unsupervised Discriminative Training with Application to Dialect Classification,” *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2444-2453, 2007.

[13] W. Shen, N. Chen, and D. Reynolds, “Dialect Recognition using Adapted Phonetic Models,” Proceedings of Interspeech, 2008.

[14] G. Choueiter, G. Zweig, and P. Nguyen,“An Empirical Study of Automatic Accent Classification,” Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2008.

[15] N. F. Chen, W. Shen, J. P. Campbell, “A Linguistically-Informative Approach to Dialect Recognition Using Dialect-Specific Context-Dependent Phonetic Models,” Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2010.

[16] N. F. Chen, W. Shen, J. P. Campbell, P. Torres-Carrasquillo, “Informative Dialect Recognition using Context-Dependent Pronunciation Modeling,” Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2011.

[17] F. Biadsy, J. Hirschberg, M. Collins, “Dialect recognition using a phone-GMM-supervector-based SVM Kernel,” Proceedings of Interspeech, 2010.

[18] M. Zissman, T. Gleason, D. Rekart, and B. Losiewicz, “Automatic Dialect Identification of Extemporaneous Conversational Latin American Spanish Speech,” Proceedings of International Conference on Acoustic, Speech, and Signal Processing, 1995.

[19] J. Navratil, “Recent advances in phonotactic language recognition using binary-decision trees,” Proceedings of Interspeech, 2006.

[20] F. Richardson and W. Campbell, “Discriminative N-Gram Selection for Dialect Recognition,” Proceedings of Interspeech, 2009.

[21] F. Biadsy, H. Soltau, L. Mangu, J. Navratil, and J. Hirschberg, “Discriminative phonotactics for dialect recognition using context-dependent phone classifiers,” Proceedings of Odyssey: The Speaker and Language Recognition Workshop, 2010.

[22] B. Bielefield, “Language identification using shifted delta cepstrum,” Proceedings of the Fourteenth Annual Speech Research Symposium, 1994.

[23] N. F. Chen, W. Shen, and J. P. Campbell, “Characterizing Deletion Transformations across Dialects using a Sophisticated Tying Mechanism,” Proceedings of Interspeech, 2011.

[24] N. F. Chen, W. Shen, and J. P. Campbell, “Analyzing and Interpreting Automatically Learned Rules Across Dialects,” Proceedings of Interspeech, 2012.

[25] E. Fosler-Lussier, “A Tutorial on Pronunciation Modeling for Large Vocabulary Speech Recognition,” in S. Renals and G. Grefenstette (Eds), *Text and Speech Triggered Info*, Springer-Verlag Berlin Heidelberg, 2003.

[26] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[27] R. Quinlan, “Induction of decision trees,” Machine Learning, vol. 1, no. 1, pp.81-106, 1986.

[28] L. Loots and T. Niesler, “Data-driven Phonetic Comparison and Conversion between South African, British and American English Pronunciations,” Proceedings of Interspeech, 2009.

[29] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, “WSJ-CAM0: A British English Corpus for Large Vocabulary Continuous Speech Recognition,” Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 1994.

[30] D. Paul and J. Baker, “The design for the Wall Street Journal-based CSR corpus,” Proceedings of DARPA Speech and Natural Language Workshop, Morgan Kaufmann, 1992.

[31] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, “WSJCAMO corpus and recording description,” Technical report: CUED/F-INFENG/TR.192, University of Cambridge, Department of Engineering, 1994.

[32] P. Kingsbury, S. Strassel, C. McLemore, and R. MacIntyre, “CALLHOME American English lexicon (Pronlex). LDC Catalog No. LDC97L20, 1997.

[33] C. Cieri, D. Miller, and K. Walker “The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text,” International Conference on Language Resources and Evaluation (LREC), 2004.

[34] N. F. Chen, W. Shen, J. P. Campbell, and R. Schwartz, “Large-Scale Analysis of Formant Frequency Estimation Variability in Conversational Telephone Speech,” Proc. of Interspeech, 2009.

[35] L. Gillick and S. Cox, “Some statistical issues in the comparison of speech recognition algorithms,” Proceedings of Acoustics, Speech, and Signal Processing, 1989.

[36] N. F. Chen, “Characterizing Phonetic Transformations and Fine-Grained Acoustic Differences across Dialects,” Ph.D. Thesis, Massachusetts Institute of Technology, 2011.

[37] StoryCorps: http://storycorps.org, last accessed: 14 November 2012.

[38] H. Giles and P. Powesland, Accommodation theory,” in: N. Coupland, A. Jaworski,(Eds.), *Sociolinguistics: A Reader and Coursebook*, Palgrave, New York, pp. 232-239, 1997.

[39] P. Trudgill, “Accommodation between dialects,” in: Linn, M.D. (Ed.), *Handbook of Dialect and Language Variation*, Academic Press, San Diego, pp. 307-342.

[40] J. L. Gauvain and C. H. Lee, “MAP Estimation of Continuous Density HMM: Theory and Applications,” DARPA Speech and Natural Language Workshop, 1992.

[41] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” Journal of Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[42] T. Kinnunen and H. Li “An Overview of Text-Independent Speaker Recognition: from Features to Supervectors,” Speech Communication, 2009.

[43] D. Reynolds, W. Campbell, W. Shen, E. Singer, “Automatic Language Recognition via Spectral and Token Based Approaches,” in J. Benesty et al. (eds.), *Springer Handbook of Speech Processing*, Springer, 2007.

[44] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” In Proc. ICSLP, 2002.

[45] W. Wolfram and N. Schilling-Estes, *American English: Dialects and Variation, 2nd Edition. Appendix: An Inventory of Socially Diagnostic Structures*, Blackwell Publishing Professional, 2005.

[46] L. Green, *African American English*, Cambridge University Press, 2002.

[47] C. Cieri, J. P. Campbell, H. Nakasone, K. Walker, and D. Miller, “The Mixer corpus of multilingual, multichannel speaker recognition data,” Pennsylvania University of Philadelphia, 2004.

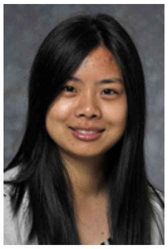
[48] J. M. Guitart, “Variability, multilectalism, and the organization of phonology in Caribbean Spanish dialects,” in F. Martinez-Gil, *Issues in the Phonology and Morphology of the Major Iberian Languages*, Georgetown University Press, 1997.

[49] J. H.-T. Yang, “The role of sound change in speech recognition research,” Conference on Computational Linguistics and Speech Processing (ROCLING). Taiwan: Academia Sinica, 2007.

[50] Personal communication with Prof. Feng Fan Hsieh at National Tsing Hua University in Taiwan, 2013.

[51] B. Yin and F. Felley, *Chinese Romanization. Pronunciation and Orthography*, Beijing: Sinolingua, 1990.

Nancy F. Chen received her Ph.D. from Massachusetts Institute of Technology (MIT) and Harvard University in 2011. She is currently a scientist at the Institute of Infocomm Research (I^2R), Singapore. Her current research interests include spoken language processing for under-resourced languages, speech summarization, spoken term detection, and computer-assisted language learning. Prior to joining I^2R, she worked at MIT Lincoln Laboratory on her Ph.D. research, which integrates speech technology and speech science, with applications in speaker, accent, and dialect characterization. Dr. Chen is a recipient of the Microsoft-sponsored IEEE Spoken Language Processing Grant, the MOE Outstanding Mentor Award, and the NIH Ruth L. Kirschstein National Research Service Award.



Sharon W. Tam received a S.B. in computer science and engineering and a M.Eng. in electrical engineering and computer science from Massachusetts Institute of Technology in 2010 and 2011, respectively. Since then, she has been a member of the Human Language Technology group at MIT Lincoln Laboratory where she is currently working on cross-language information retrieval.

Wade Shen Mr. Wade Shen is currently an Assistant Group Leader in the Human Language Technology Group at MIT Lincoln Laboratory. His current areas of research involve machine translation and machine translation evaluation; speech, speaker, and language recognition for small-scale and embedded applications; named-entity extraction; and prosodic modeling.

Prior to joining Lincoln Laboratory in 2003, Mr. Shen helped found and served as Chief Technology Officer for Vocentric Corporation, a company specializing in speech technologies for small devices.

Mr. Shen received his Master’s degree in computer science from the University of Maryland, College Park, in 1997, and his Bachelor’s degree in electrical engineering and computer science from the University of California, Berkeley, in 1994.

Joseph P. Campbell (S '90, M '92, SM '97, F '05) received B.S., M.S., and Ph.D. degrees in electrical engineering from Rensselaer Polytechnic Institute in 1979, The Johns Hopkins University in 1986, and Oklahoma State University in 1992, respectively. Dr. Campbell is currently the Associate Group Leader of the Human Language Technology Group at MIT Lincoln Laboratory, where he directs the group's research in speech, speaker, language, and dialect recognition; word and topic spotting; speech and audio enhancement; speech coding; text processing; natural language processing; machine translation of speech and text; information retrieval; extraction of entities, links and events; multimedia recognition techniques, including both voice and face recognition for biometrics applications; and advanced analytics for analyzing social networks based on speech, text, video, and network communications and activities. Joe specializes in research, development, evaluation, and transfer of speaker recognition technologies, evaluation and corpus design, and also specializes in biometrics for government applications. He chairs the IEEE Jack S. Kilby Signal Processing Medal Committee and the International Speech Communication Association's Speaker and Language Characterization Special Interest Group (ISCA SpLC SIG). Dr. Campbell is a member of the IEEE Medals Council, the International Speech Communication Association, Sigma Xi, the Boston Audio Society, and the Acoustical Society of America. Dr. Campbell was named a Fellow of the IEEE "for leadership in biometrics, speech systems, and government applications" in 2005.

Before joining Lincoln Laboratory as a Senior Staff member in 2001, Dr. Campbell worked for the US Government (1979-2001); chaired the Biometric Consortium (1994-1998); taught Speech Processing at The Johns Hopkins University (1991-2001); and was an Associate Editor of the *IEEE Transactions on Speech and Audio Processing* (1991-1999), an IEEE Signal Processing Society Distinguished Lecturer (2001-2002), a member of the IEEE Signal Processing Society's Board of Governors (2002-2004), a coeditor of *Digital Signal Processing* journal (1998-2005), a member of the IEEE Information Forensics Security Technical Committee (2005-2009), the Vice President of Technical Activities of the IEEE Biometrics Council (2008-2011), and a member of the National Academy of Sciences' *Whither Biometrics?* Committee that produced the book *Biometric Recognition: Challenges and Opportunities* (2004-2010).