Analyzing and Interpreting Automatically Learned Rules Across Dialects*

Nancy F. Chen$^{1,2}$, Wade Shen$^1$, Joseph P. Campbell$^1$

$^1$MIT/Lincoln Laboratory, Lexington, MA, USA
$^2$Institute for Infocomm Research, Singapore, Singapore

nancycheden@alum.mit.edu, swade@ll.mit.edu, jpc@ll.mit.edu

Abstract

In this paper, we demonstrate how informative dialect recognition systems such as acoustic pronunciation model (APM) help speech scientists locate and analyze phonetic rules efficiently. In particular, we analyze dialect-specific characteristics automatically learned from APM across two American English dialects. We show that unsupervised rule retrieval performs similarly to supervised retrieval, indicating that APM is useful in practical applications, where word transcriptions are often unavailable. We also demonstrate that the top-ranking rules learned from APM generally correspond to the linguistic literature, and can even pinpoint potential research directions to refine existing knowledge. Thus, the APM system can help phoneticians analyze rules efficiently by characterizing large amounts of data to postulate rule candidates, so they can reserve time to conduct more targeted investigations. Potential applications of informative dialect recognition systems include forensic phonetics and diagnosis of spoken language disorders.

Index Terms: informative dialect recognition, rule retrieval, phonological rules, forensic phonetics

1. Introduction

While many dialect identification (DID) systems take advantage of acoustic or phonotactic differences, most of them do not focus on explicitly characterizing the dialect differences in terms of rules. A few exceptions include [1, 2, 3, 4, 5]. In [1, 2], discriminative classifiers were trained to recognize dialects, and N-grams or triphones helpful in DID were discussed. In [3], we modeled phonetic transformations across dialects, and explicitly characterized them into insertions, substitutions, and deletions using phonetic pronunciation model (PPM), depicted in Fig. 1. In [4], acoustic differences caused by phonetic context were used to infer underlying phonetic rules. In [5], we integrated [3] and [4] and further proposed an acoustic counterpart of PPM and used clustering to generalize phonetic context containing dialect-specific information. We term this line of work explicitly characterized them into insertions, substitutions, and deletions using phonetic pronunciation model (PPM), depicted in Fig. 1. In [4], acoustic differences caused by phonetic context were used to infer underlying phonetic rules. In [5], we integrated [3] and [4] and further proposed an acoustic counterpart of PPM and used clustering to generalize phonetic context containing dialect-specific information. We term this line of work phonotactic differences, most of them do not focus on explicitly characterizing the dialect differences in terms of rules. A few exceptions include [1, 2, 3, 4, 5]. In [1, 2], discriminative classifiers were trained to recognize dialects, and N-grams or triphones helpful in DID were discussed. In [3], we modeled phonetic transformations across dialects, and explicitly characterized them into insertions, substitutions, and deletions using phonetic pronunciation model (PPM), depicted in Fig. 1. In [4], acoustic differences caused by phonetic context were used to infer underlying phonetic rules. In [5], we integrated [3] and [4] and further proposed an acoustic counterpart of PPM and used clustering to generalize phonetic context containing dialect-specific information. We term this line of work informative dialect recognition, rule retrieval, phonological rules, forensic phonetics

*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.

Nancy Chen performed part of the work while a PhD student at MIT / Lincoln Laboratory.

Table 1: Data partition of the StoryCorps corpus.

<table>
<thead>
<tr>
<th>Set</th>
<th>Duration (hr)</th>
<th># of Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>training</td>
<td>22.6</td>
<td>69</td>
</tr>
<tr>
<td>developmental</td>
<td>7.1</td>
<td>38</td>
</tr>
<tr>
<td>test</td>
<td>7</td>
<td>28</td>
</tr>
</tbody>
</table>

2. Material and Model

2.1. Corpus

StoryCorps [8] includes 2 American English dialects: (1) African American Vernacular English (AAVE) speakers who self-reported as black or African Americans; (2) Non-AAVE speakers who self-reported as white. All speech data chosen are conversations between friends or family members that speak the same dialect to minimize accommodation issues [9]. In addition, since gender and age also affect dialects, these two factors were also balanced across all data sets and dialect groups. The data partition of the corpus is shown in Table 1.

2.2. Pronunciation Model

Suppose we are given a dataset consisting of speech utterances (e.g., AAVE), their corresponding word transcriptions and a reference pronunciation dictionary (e.g., non-AAVE). For each speech utterance, we can generate reference phones $C = c_1, c_2, ..., c_n$, obtained through using the word transcriptions and pronunciation dictionary and surface phones $O = o_1, o_2, ..., o_T$, obtained through phone decoding of the AAVE speech utterances.

A hidden Markov model (HMM) can be used to model the relationship between the reference phones and surface phones; i.e., how the phones of the reference dialect (e.g., non-AAVE) transforms to the phones of the dialect of interest (e.g., AAVE). These phonetic transformations are categorized as substitution, insertion, and deletion rules. The traditional HMM network is not suitable for modeling insertion and deletion rules explicitly, so phonetic pronunciation model (PPM) was proposed in [3] to address these issues. The network of PPM is shown in...
3. Dialect Identification Experiment

In this experiment, we examine the fusion gains of the APM system with other dialect identification systems, and discuss the implications of the results.

3.1. Implementation Details

The developmental and test set were segmented into 30-sec trials to perform likelihood ratio tests. The developmental set was used to train a Gaussian backend classifier [10] used for the fusion experiments.

3.1.1. Acoustic Pronunciation Model (APM)

Reference phones were obtained through forced alignment using the word transcripts and a root phone recognizer trained on the corpus WSJ0 [11]. The front-end features are perceptual linear predictive coefficients [12], 1st delta features, and 2nd delta features, and are modeled as Gaussian mixtures. We first train a monophone HMM system, and its model parameters are used to initialize a triphone system. Then, decision tree clustering [13] was used to tie triphone states and learn dialect-specific rules [7]. No word transcriptions were used during test time.

3.1.2. Adapted phonetic models (APM0)

Adapted phonetic models (denoted as APM0 in this work) is a monophone-based system proposed in [18]. The root phone recognizer used in Section 3.1.1 was used to decode the training data into APM0’s reference phone units. The acoustic observations corresponding to these reference phone units were modeled as Gaussian mixtures. All other experimental setup is the same as APM in Section 3.1.1.

3.1.3. Gaussian Mixture Model using Shifted Delta Cepstrum (GMM-SDC)

Shifted delta cepstrum (SDC) are used as front-end features. Each GMM has 2048 mixture components. A universal background model (UBM) was first trained on the training data of all dialects, and then the means of the dialect-specific GMMs were adapted using Maximum A Posteriori (MAP) [14]. All experimental setup is the same as [15].

3.1.4. Parallel Phone Recognition followed by Language Modeling (PPRLM)

The adapted phonetic models trained in Section 3.1.2 were used as tokenizers to generate phonetic sequences to train trigram language models, as in [18].

3.2. Results and Discussion

Table 2 shows the equal error rates (EER) of the dialect identification tasks. We see that the acoustic systems APM, APM0, and GMM-SDC perform similarly, all close to 11%. The phonetic system PPRLM performs slightly worse; the EER is around 13.5%.

Table 3 shows the dialect identification results after APM0, GMM-SDC, and PPRLM were each individually fused with APM. The fused EERs are all reduced compared to the original systems. PPRLM’s relative fusion gain is the greatest, reaching 46%; those of APM0 and GMM are around 25%. These fusion gains suggest that APM models information complementary to the other systems.

We speculate that this complementary information is from APM’s explicit characterization of dialect rules. During training, the use of reference phones derived from the word transcript provides a canonical and clean basis for characterizing dialect differences. The network architecture of APM also provides explicit modeling of dialect-specific rules, while standard systems only do so implicitly. In the next experiment, we further analyze how well APM retrieves rules, and compare the learned rules with the literature.

Table 2: Dialect identification results.

<table>
<thead>
<tr>
<th>System</th>
<th>Equal Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>APM</td>
<td>10.73</td>
</tr>
<tr>
<td>APM0</td>
<td>10.76</td>
</tr>
<tr>
<td>GMM-SDC</td>
<td>10.86</td>
</tr>
<tr>
<td>PPRLM</td>
<td>13.45</td>
</tr>
</tbody>
</table>

Table 3: Fusion results with APM.

<table>
<thead>
<tr>
<th>Fused System</th>
<th>Equal Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>APM0 + APM</td>
<td>8.1</td>
</tr>
<tr>
<td>GMM-SDC + APM</td>
<td>7.9</td>
</tr>
<tr>
<td>PPRLM + APM</td>
<td>7.2</td>
</tr>
</tbody>
</table>
4. Rule Retrieval Experiment

4.1. Experimental Setup

4.1.1. Ground-truth rules

A group of phoneticians extracted literature descriptions of the AAVE dialect (e.g., [16, 17]) and converted them to phonetic rules. These rules are in the 1st column of Fig. 3, and we adopt them as ground-truth. The occurrence frequency of these rules in the training set are in the 2nd column of Fig. 3.

4.1.2. True trials vs. false trials

The reference phones were obtained by force-alignment using a phone recognizer trained on non-AAVE speech with the WSJ0 pronunciation dictionary. The surface phones were obtained similarly using a phone recognizer trained on AAVE speech with an AAVE version of the WSJ0 pronunciation dictionary: for each pronunciation in the dictionary, an AAVE alternative pronunciation was generated according to the ground-truth rules. The most likely alignment between the reference phones and surface phones were obtained by the Viterbi algorithm.

A true trial is a reference triphone occurrence that matches the ground-truth rule’s phone of interest, but either (a) the reference triphone’s phonetic context does not match the rule, or (b) the reference triphone’s corresponding surface phones does not match the transformed phone in the rule.

A false trial is a reference triphone occurrence that matches the ground truth rule’s phone of interest, but both (a) the reference triphone’s phonetic context does not match the rule, and (b) the surface phone corresponding to the reference triphone matches the transformed phone in the rule.

4.1.3. Scoring

We used supervised scoring (word transcriptions used to generate reference phones during test time) as an upper-bound for performance evaluation. The duration-normalized log likelihood ratio (LLR) of each test trial was used to compute recall and precision rates. We use the F-measure, harmonic mean of recall and precision, to represent retrieval performance.

We used supervised scoring (word transcriptions used to generate reference phones during test time) as an upper-bound for performance evaluation. The duration-normalized log likelihood ratio (LLR) of each test trial was used to compute recall and precision rates. We use the F-measure, harmonic mean of recall and precision, to represent retrieval performance.

Let \( O_i \) be an acoustic observation of test trial \( i \) and \( T_i \) be the duration of \( O_i \). LLLR of \( O_i \) is

\[
\frac{1}{T_i} \log \frac{P(O_i|\lambda_{d_1}, C_i)}{P(O_i|\lambda_{d_2}, C_i)}
\]

(1)

where \( C_i \) is the corresponding reference phones of \( O_i \). \( \lambda_{d_1} \) and \( \lambda_{d_2} \) are APM models of dialect 1 (i.e., AAVE) and dialect 2 (i.e., non-AAVE).

Unsupervised rule retrieval experiments were conducted similarly to the supervised case, except that LLR was computed without knowing the reference phones at test time. Instead, observations corresponding to the time regions matching those of the true and false trials in the supervised case were used to compute LLLR. Therefore, for test trial \( i \), Eq. (1) becomes:

\[
\frac{1}{T_i} \log \frac{P(O_i|\lambda_{d_1})}{P(O_i|\lambda_{d_2})}
\]

(2)

4.2. Hypothesis

We hypothesize that unsupervised rule retrieval performance is close to the supervised case. The logic behind is that while we do not know what the reference phones are nor where they are located in the unsupervised case, we do know where large LLRs between the non-AAVE and AAVE models are located, implying the locations of dialect-specific information.

4.3. Results

As shown in Fig. 2, the unsupervised retrieval results are close to the supervised retrieval results, supporting our hypothesis. The weighted average retrieval performance of the supervised and unsupervised care are similar: 47.27% and 45.02%, respectively. (Weights were determined by how prevalent the center phone of the triphone is in the training set). Therefore, although word transcriptions are required during training, no word transcripts are needed during test time. This is useful in practical scenarios (e.g., forensic applications), where word transcriptions are often unavailable.

4.4. Further analysis

In this section, we compare the top ranking rules learned from APM to the linguistics literature.

4.4.1. Ranking learned rules

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.

4.4.2. Comparison with linguistic literature

In this section, we compare the top ranking rules learned from APM to the linguistics literature.
coding consonants, the system suggests that [ay] only deglides when preceding certain consonants (e.g., [w], [m], [hh]) Fig. 4 shows a spectrogram example, where degliding only occurs in one out of three instances of [ay] followed by a consonant.

For the rule of 1 vocalization ([l] —> / [+vowel] — [+const]), the corresponding learned rule has a more specified left-context but less specified right-context; the learned rule specifies the left-context to be a high, back vowel (i.e., [uw]), while the right-context is unspecified. Discrepancies like these could be used as working hypotheses to help phoneticians further investigate whether the existing rules need to be refined.

The APM system can help phoneticians analyze rules more efficiently by characterizing large amounts of data to postulate hypothesized rules, saving phoneticians time and manual effort. With the postulated rules from the system, phoneticians can conduct more targeted and detailed investigations to either refine/verify existing rules or potentially propose new rules.

4.5. Discussion: Potential Applications

- **Forensic Phonetics:** In forensic phonetics, it is important to tease apart which features are specific to a particular speaker and which features are specific to the dialect(s) the speaker belongs to [19]. However, the dearth of dialect studies on large corpora makes it extremely challenging to establish quantifiable norms for dialect features. Our proposed system can resolve this issue as it is able to characterize large-corpora very efficiently.

- **Speech and Language Pathology:** The framework of informative dialect recognition and the APM approach in not limited to only analyzing dialect differences; it has clinical implications as well. For example, speech and language pathologists often have limited time to interact with the client, making it difficult to analyze sufficient data to determine the suitable treatment. The APM system could solve this problem by characterizing large amounts of off-site recorded data. This procedure helps the clinicians make more targeted diagnoses and measure treatment progress more effectively and efficiently.

5. Conclusions

In this paper, we demonstrated how informative dialect recognition systems (e.g., APM) can help speech scientists analyze phonetic rules efficiently. In particular, we examined dialect-specific rules automatically learned from APM across two American English dialects. Dialect identification experiments showed that APM fuses well with other standard systems, suggesting that APM exploits complementary information from explicit characterization of rules not modeled in other systems. Rule retrieval experiments demonstrated that unsupervised rule retrieval performs similarly to supervised retrieval, indicating that APM is useful in practical scenarios (e.g., forensic applications), where word transcripts are often unavailable. In addition, top-ranking rules learned from APM generally correspond to the linguistic literature; some even provide insight into how currently known dialect-specific rules could be refined. Potential applications of informative dialect recognition systems include assistive diagnosis tools for spoken language disorders.

6. References


