Discriminative pronunciation modeling for dialectal speech recognition

Maider Lehr[1], *Kyle Gorman*[1], *Izhak Shafran*[2]

[1]Center for Spoken Language Understanding, OHSU, Portland, OR, USA
[2]Google Inc

lehrm,gormanky@ohsu.edu, izhak@google.com

Abstract

Speech recognizers are typically trained with data from a standard dialect and do not generalize to non-standard dialects. Mismatch mainly occurs in the acoustic realization of words, which is represented by acoustic models and pronunciation lexicon. Standard techniques for addressing this mismatch are generative in nature and include acoustic model adaptation and expansion of lexicon with pronunciation variants, both of which have limited effectiveness. We present a discriminative pronunciation model whose parameters are learned jointly with parameters from the language models. We tease apart the gains from modeling the transitions of canonical phones, the transduction from surface to canonical phones, and the language model. We report experiments on African American Vernacular English (AAVE) using NPR's StoryCorps corpus. Our models improve the performance over the baseline by about 2.1% on AAVE, of which 0.6% can be attributed to the pronunciation model. The model learns the most relevant phonetic transformations for AAVE speech.

Index Terms: large vocabulary speech recognition, dialectal speech recognition, pronunciation modeling, discriminative training

1. Introduction

Speech recognition technology is increasingly ubiquitous in everyday life. Automatic speech recognition (ASR) is used to interact with customer service systems and personal electronic devices. Medical professionals use ASR for dictation, and clinicians and educators employ it for automated assessment [1, 2, 3]. Differences between individual speakers pose one of the main challenges in speech recognition. Speech varies with age, gender, ethnicity, geography (in the form of regional dialects), and socioeconomic status [4, 5]. In addition to this between-speaker variation, speech "style" also impacts recognition quality. Features of casual speech—greater speech rate and higher rates of fillers (*uh* and *um*), repetitions, false starts, and vocal noise—make recognition more difficult than for formal styles like those used in news broadcasts. However, developing new ASR systems, for example, for specific dialects requires a significant investment in preparing the necessary corpus of manually transcribed speech data (with an accompanying pronunciation lexicon).

In traditional ASR systems, pronunciation is represented by context-dependent (CD) acoustic states at acoustic level, and by the pronunciation lexicon at phonetic level. The pronunciation lexicon is usually a simple, deterministic mapping from phones to words using canonical pronunciations. Traditionally, both the acoustic model (AM) and pronunciation lexicon are independently adapted when extending an ASR system to new dialects. Acoustic models are adapted using techniques such as generative maximum a posteriori (MAP) [6] or maximum likelihood linear regression (MLLR) techniques [7, 8, 9, 10], and in some cases new pronunciation variants are added to the pronunciation lexicon [11, 12, 13, 14, 15, 16]. Another alternative is to modify the context-dependent decision trees [9, 17]. In addition, when dealing with dialectal speech recognition the language model could also be adapted, although it is generally assumed that the degradation due to grammatical mismatches between two dialects is less than that caused by pronunciation-related mismatches.

One common approach to learn new pronunciations employs joint-multigram models which map graphemes onto phones [18]. They are built as regular word n-gram language models (LMs), but the tokens are phone/grapheme pairs—graphone units—instead of words. These models capture information about phone and grapheme context [19, 20]. Alternatively, phonological rules can be represented with the joint multigram model. In this case the model maps spoken phone sequences onto canonical phone sequences [21].

Jyothi and colleagues [22] discriminatively estimate parameters from the lexicon using a weighted finite state transducer (WFST) framework. The recognizer is represented as WFST factors or components, and they learn the parameters of the arcs from the pronunciation lexicon WFST in isolation, or together with parameters from the acoustic model and context dependent decision tree WFSTs.

In this study, we learn a discriminative linear model to improve the pronunciation model of a baseline broadcast news recognizer for recognizing AAVE dialect. Continuing our thread of previous work on using discriminative linear models [23], we learn the relevant phonetic transformations for the dialect from the data jointly with the parameters from the language model.

1.1. African American Vernacular English

African American Vernacular English (AAVE) is a dialect of English spoken by most African Americans, particularly younger, urban, and working-class individuals, and particularly in casual speech [24]. AAVE is distinct from Standard American English (SAE) in grammar, pronunciation, and vocabulary. Table 1, based on Rickford (1999), lists a number of "variable rules" of AAVE. These relate canonical SAE pronunciations to common realizations in AAVE. These rules apply stochastically rather than categorically, and probability of application is sensitive to phonological context as well as speaker-specific social factors such as socioeconomic status [25, 26]. Some of these variable rules are also found in other English dialects.

1.2. Discriminative pronunciation variation model

The performance of an ASR system often degrades considerably when it is employed to recognize certain dialects that are

Phonological rule	Example	Phonological rule	Example
th → t	*thin*	[+cons] → ∅ / [+cons] __ |	*hand*
dh → d	*this*	s p → p s / __ |	*grasp*
th → f	*bath*	s t → t s / __ |	*pianist*
dh → v	*with*	s k → k s / __ |	*ask*
oy → ao	*boy*	ax → ∅ / | __	*about*
ay → aa	*right*	ih → ae / __ ng	*drink*
v → b / __ [+nasal]	*movement*	eh → ih / __ [+nasal] {|, [+cons]}	*pin*
l → ax / [+vowel] __	*bell*	l → ∅ / [+vowel] __ [+labial]	*help*
z → d / __ [+nasal]	*isn't*	z → d / [+vowel] __ [+vowel]	*reason*
r → ax / [+vowel] __	*four*	v → b / [+vowel] __ [+vowel]	*having*
r → ∅ / [+vowel] __	*there*	d → t / [+vowel] __ |	*god*
ng → n / ih __ |	*walking*	b → p / [+vowel] __ |	*cab*
r → ∅ / | th __	*throw*	t → k / s __ r	*destroy*
ts → ih z / s __ |	*breasts*	g → k / [+vowel] __ |	*log*
ts → ∅ / __ |	*ghosts*		

Table 1: Linguistic rules mapping from SAE pronunciations to possible AAVE pronunciations.

not well represented in the training data. We assume that pronunciation variation is the primary cause of this degradation. The pronunciation lexicon is traditionally a deterministic mapping from words into their canonical phonetic representation. There may be a few pronunciation variants for some words (e.g., *and*). The lexicon is not learned or adapted for each task, so pronunciation variation is largely captured by the AM [27]. However, it is not often feasible to train new AMs for other dialects, so baseline AMs, trained on publicly available corpora, must be adapted for recognition of those other dialects. Furthermore, when dealing with speech from a dialect other than the standard the pronunciation at the lexical level differs more from the canonical pronunciations of the lexicon. Hence, the lexical-level pronunciation must be adapted too.

The adaptation may be performed by adding new pronunciation variants into the lexicon, or by learning a pronunciation variation model that maps surface phones onto canonical phones. The pronunciation variation model may contain knowledge-based rules or rules learned from the data. In our work we use knowledge-based phonological rules, depicted in Table 1, to generate the features for the discriminative pronunciation variation model. Specifically, we discriminatively learn the weights for n-grams of phone transformations of the form $p_s : p_c$ that match any rule from Table 1, where p_s is the surface phone and p_c the canonical phone. We apply the idea of the global linear models [28] for the estimation of the discriminative models with the perceptron algorithm, as in [23].

Briefly, the goal of the decoding is to estimate the word sequence w for a given input speech utterance x. For that, the speech utterances x are decoded to get the ASR outputs y. A function **GEN** enumerates a set of candidates **GEN**(x) for the input x (for example, N-best candidates or lattices). In our case, each candidate y does not only contain the word sequence, but also a sequence with phonetic information, where each token is a phone transformation $p_s : p_c$. In the proposed discriminative pronunciation variation model, the representation Φ maps each (x, y) to a feature vector $\Phi_p(x, y) \in \Re^d$ and a parameter vector $\bar{\alpha}_p \in \Re^d$ (features and parameters related to the pronunciation variation model are denoted with the subscript p), and the output of a linear model $F(x)$ is computed as below.

$$F(x) = \underset{y \in \mathbf{GEN}(x)}{\arg\min}\ \Phi_p(x, y)\bar{\alpha}_p$$

2. Experimental results

2.1. Corpus

We use data from the StoryCorps project[1] for all experiments, a subset from the data set used by Chen and colleagues for their dialect recognition experiments [29]. This data consists of conversations between two or more speakers of the same dialect, either AAVE or SAE. Speaker dialect is self-reported. We learn independent discriminative models for each dialect. For AAVE, the training and test set contain 56 speakers and 11 speakers, respectively. The training and test set for SAE include 46 speakers and 14 speakers, respectively. For both dialects, the training set consists of approximately 13 hours of speech, and the test set of approximately 3 hours. Both training and test sets are balanced in terms of female and male speech quantity.

2.2. Baseline system

The baseline ASR system is a recognizer designed for broadcast news modeled with the IBM toolkit in a similar way as presented in the work by Soltau and colleagues [30]. The acoustic model consists of 4000 clustered allophone or acoustic states defined over a pentaphone context and a set of 44 phones, with states represented by Gaussian mixture models with a total of 150K mixture components. The observation vectors consist of PLP features, stacked from 10 neighboring frames and projected into a 50-dimension space using linear discriminant analysis. The acoustic models were trained using 430 hours of transcribed broadcast news speech [31, 32]. The language model was estimated using several corpora of conversational telephone speech; it has a 48k-word vocabulary and contains approximately 16M, 16M, and 11M bigrams, trigrams and 4-grams, respectively. On average, the pronunciation lexicon has 1.1 pronunciations per word. Decoding is performed in several stages using successively refined acoustic models, including a context-dependent model, a vocal-tract normalized model, a speaker-adapted maximum likelihood linear regression model, and a feature and model-based discriminative model [33].

[1] www.npr.org/series/4516989/storycorps

2.3. Generation of competing candidates

For training the dialectal linear discriminative models—the AAVE and SAE models—each utterance x is decoded with the baseline recognizer, generating lattices L_x containing word sequences and associated log probabilities. The 100-best unique candidates are then extracted from each lattice. For each candidate y, apart from the word sequence we also include the canonical phone sequence p_c. The canonical phone sequence is obtained by performing a forced-alignment of the speech utterance with the word sequence candidate using the baseline (out-of-dialect) ASR.

We generate a second phone sequence for each candidate using the knowledge-based phonological rules from Table 1. For each word we include the canonical and new additional pronunciations generated after applying the phonological rules to the canonical pronunciation. Table 2 shows the generated lexicon for a candidate with the word sequence "*and then*". In this example, a new pronunciation is generated for the word *and* after applying the rule [+cons] $\rightarrow \emptyset$. The word *then* gets assigned two new pronunciations from the rule dh $\rightarrow$ d and dh $\rightarrow$ v.

and(01)	ae	n	d
and(02)	ae	n	
then(01)	dh	eh	n
then(02)	d	eh	n
then(03)	v	eh	n

Table 2: Extended lexicon for candidate "*and then*".

A forced-alignment of the word sequence candidate is performed with the extended lexicon. The phone sequence associated with the lowest acoustic score may be different from the canonical phone sequence. We call this new phone sequence the surface phone sequence p_s. Then, the Levenshtein alignment is computed for the canonical phone p_c and surface phone sequence p_s. The alignment produces a sequence whose tokens are pairs of phones of the form $p_s : p_c$, where p_s is the surface phone and p_c the canonical phone. These pairs are useful for incorporating weights for specific transformations in the discriminative linear model.

2.4. Parameter estimation

The parameters of the models are iteratively estimated with perceptron algorithm, using the oracle candidates as the reference. The parameter estimation is iterated until we see no improvement on the held-out data for five iterations. Maximum likelihood scores, L_{ML}, are interpolated with scores from the discriminative models, L_{DM}, using an interpolation weight α_0 determined by 20-fold cross validation.

$$L = \alpha_0 * L_{ML} + (1 - \alpha_0) * L_{DM}$$

2.5. Feature Space

For each candidate in the N-best list we have the following sequences:

- Word sequence: <s> and then < /s>
- Canonical phone sequence: | ae n d | th eh n|
- Surface phone sequence | ae n | t eh n |
- Joint phone sequence: | ae:ae n:n -:d | t:th eh:eh n:n |

2.6. Results

2.6.1. Discriminative pronunciation variation model

We compare our models on StoryCorps corpus, using both the AAVE and SAE portions described in Section 2.1 and report Word Error Rate (WER) on 20-fold cross-validation (Xval) and held-out test set. As explained earlier, the pronunciation-related features are unigrams and bigrams of phone transformations; $\Phi_{p_s:p_c}$. The experiments on SAE and AAVE were performed independently to understand the effectiveness of the model on standard vs dialectal speech. Statistical significance was computed with respect to the baseline system using the matched-pairs test [34] provided by NIST SCTK toolkit [35] and non-significant results at p<0.001 are marked by †.

All experiments were performed using baseline discriminative acoustic models (BMMI) described in Section 2.2 and the results are reported in Table 3.

The performance of the baseline ASR models on AAVE is over 10% worse than on SAE. This is expected since the SAE is closer to the training data of the baseline acoustic models (Broadcast News). For reference, we report the oracle accuracy for the 100-best candidates which specifies the lowest achievable WER for all the re-scoring experiments reported in the table. There is considerable room for improvement in both SAE and AAVE sections.

Next, we re-score the N-best candidates with the baseline acoustic models and the extended lexicon that incorporates the phonological rules from Table 1. The performance of this maximum likelihood re-scoring pass is denoted as "ML re-scoring". The WER improves on the AAVE portion but not on the SAE portion. This may be an indicator of the utility and relevance of the phonological rules for AAVE.

Instead of ML re-scoring, we use canonical phones Φ_c as features in a discriminative model. The performance of this model is marginally better than the ML re-scoring for both SAE and AAVE. Increasing the complexity of the model, we investigate the utility of features that encode both the transformed surface form and canonical forms $\Phi_{p_s:p_c}$. The additional complexity does not appear to be particularly useful in this data set. The learned transforms may be useful in identifying a smaller subset of useful transforms for ML re-scoring, but we did not explore this further.

2.6.2. Joint discriminative pronunciation and language models

As the proposed estimation framework allows us to easily learn parameters from different ASR components jointly, we analyze the interaction between the language model and the pronunciation model. The language model is represented using unigram and bigrams word features Φ_w.

As shown in Table 3 discriminative language models (DLMs) are more useful for AAVE data than for SAE data. They provide gains about 1.5% for the AAVE test set, but about 0.5% for the SAE test set.

Next, we investigate the effect of jointly estimating the parameters of the discriminative language model Φ_w and the pronunciation model, either with Φ_{p_c} or with $\Phi_{p_s:p_c}$. Including pronunciation features into the discriminative language models provides further consistent gains, especially for the AAVE data. The best results among all models are obtained when the discriminative language model and the discriminative pronunciation models are estimated jointly.

In summary, a discriminative model that jointly estimates the parameters of the language model and the pronunciation

	SAE					AAVE				
	$\#F_U$	α_0	Xval	$\#F_C$	test	$\#F_U$	α_0	Xval	$\#F_C$	test
Baseline	-	-	22.3	-	25.2	-	-	36.1	-	38.2
Oracle	-	-	14.8	-	16.6	-	-	26.7	-	28.1
Pronunciation model										
ML re-scoring	-	-	22.4†	-	25.3†	-	-	35.9	-	37.7
Φ_{p_c}	0.9K	0.45	21.8	18.5M	25.0†	1.0K	0.35	35.1	20.7M	37.0
$\Phi_{p_s:p_c}$	1.4K	0.45	21.9	18.3M	25.1†	1.5K	0.35	35.1	20.7M	36.9
Joint pronunciation model										
Φ_w	48.5K	0.25	21.8	6.9M	24.7	61.0K	0.25	34.9	4.1M	36.7
Φ_w, Φ_{p_c}	60.5K	0.45	21.5	22.0M	**24.4**	76.8K	0.40	34.6	24.7M	36.2
$\Phi_w, \Phi_{p_s:p_c}$	59.9K	0.55	21.7	21.9M	24.5	67.1K	0.45	34.6	24.8M	**36.1**

Table 3: Comparison of performance (WER) of discriminative models on SAE and AAVE portions of the corpus using features from different levels—words (Φ_w), canonical forms (Φ_{p_c}) and associated transformation of canonical form ($\Phi_{p_s:p_c}$). F_u and F_c denote the number of unique features in the model and the number of times they were employed in the respective test sets, respectively.

model improves the performance of the AAVE recognizer by 2.1% WER of which 0.6% can be attributed to pronunciation models. As expected, improvements on the SAE data are smaller since the canonical pronunciations are well-matched to SAE and the training data used in the baseline recognizer.

3. Summary

In this paper, we develop a discriminative pronunciation model which can be jointly estimated with the discriminative language model. We evaluate the effectiveness of the models on a corpus of SAE and AAVE speakers. Our results show that the discriminative language model and the discriminative pronunciation model improve the performance of the AAVE data significantly more than of the SAE data. The discriminative pronunciation model is more effective than an equivalent ML model. We also obtain further gains when both the language model and the pronunciation model are jointly estimated. In all, we report a gain of 0.8% WER on SAE and 2.1% WER on AAVE.

When the knowledge-based rules are not available, the phone transformations could be extracted from the training data, for example, by running Levenshtein alignment between the 1-best phone sequence and reference phone sequence for each training sample. Then, we could learn the weights for the data-driven phone transformations with the proposed discriminative pronunciation model. The discriminative model should filter out the transformations that are not relevant and just assign weights to the transformations that characterize the dialect from the task data.

In future publication, we will report results on our ongoing work where we find that, despite their simplicity, the allophone state transitions provide most of the gain observed from the phone-based pronunciation models and that the discriminative models provide gains even after acoustic model adaptation.

4. Acknowledgements

We thank NPR's StoryCorps for providing us with the data used in this work; and Brian Kingsbury and IBM for the use of their ASR software tools. This research was supported by Google, Intel and IBM awards as well as NSF awards IIS 0964102 and 1027834, and NIH award K25 AG033723. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not reflect the views of the funding agencies.

5. References

[1] C. Cucchiarini, H. Strik, and L. Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[2] M. Lehr, E. Prud'hommeaux, I. Shafran, and B. Roark, "Fully automated neuropsychological assessment for detecting Mild Cognitive Impairment," in *Proceedings of Interspeech*, 2012.

[3] M. Lehr, I. Shafran, E. Prud'hommeaux, and B. Roark, "Discriminative joint modeling of lexical variation and acoustic confusion for automated narrative retelling assessment," in *Proceedings of NAACL*, 2013, pp. 211–220.

[4] W. Labov, *Principles of linguistic change: Social factors*. Malden, MA: Wiley-Blackwell, 2001.

[5] ——, *The social stratification of English in New York City*, 2nd ed. Cambridge: Cambridge University Press, 2006.

[6] J.-L. Gauvain and C. hui Lee, "Map estimation of continuous density hmm: theory and applications," in *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992, pp. 185–190.

[7] C. J. Legetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hmm," *Computer Speech and Language*, pp. 9:171–185, 1995.

[8] L. Tomokiyo and A. Waibel, "Adaptation methods for non-native speech," in *Proceedings of Multilinguality in Spoken Language Processing*, 2001.

[9] Z. Wang, T. Schultz, and A. Waibel, "Comparison of aocustic model adaptation techniques on non-native speech," in *Proceedings of ICASSP*, 2003.

[10] Y. Zheng, R. Sproat, L. Gu, I. Shafran, H. Zhou, Y. Su, D. Jurafsky, R. Starr, and S.-Y. Yoon, "Accent detection and speech recognition for shanghai-accented mandarin," in *Proceedings of Interspeech*, 2005.

[11] J. Humphries and P. C. Woodland, "Using accent-specific pronunciation modelling for improved large vocabulary continuous speech recognition," in *Proceedings of Eurospeech*, 1997.

[12] S. Goronzy, S. Rapp, and R. Kompe, "Generating non-native pronunciation variants for lexicon adaptation," *Speech Communication*, vol. 42(1), pp. 109–123, 2004.

[13] L. M. Tomokiyo, "Lexical and acoustic modeling of non-native speech in LVSCR," in *Proceedings of Interspeech*, 2000, pp. 346–349.

[14] K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," in *Proceedings of ICASSP*, 2000, pp. 1842–1845.

[15] O. Vinyals, L. Deng, D. Yu, and A. Acero, "Discriminative pronunciation learning using phonetic decoder and minimum-classficiation-error criterion," in *Proceedings of ICASSP*, 2009.

[16] B. Hutchinson and J. Droppo, "Learning non-parametric models of pronunciation," in *Proceedings of ICASSP*, 2011, pp. 4904–4907.

[17] U. Nallasamy, F. Metze, and T. Schultz, "Enhanced polyphone decision tree adaptation for accented speech recognition," in *Proceedings of Interspeech*, 2012.

[18] M. Bisani and H. Ney, "Investigations on joint-multigram models for grapheme-to-phoneme conversion," in *In Proceedings of Int. Conf. on Spoken Language Processing*, 2010, pp. 105–108.

[19] I. McGraw, I. Badr, and J. R. Glass, "Learning lexicons from speech using a pronunciation mixture model," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, pp. 357–366, 2013.

[20] L. Lu, A. Ghoshal, and S. Renals, "Acoustic data-driven pronunciation lexicon for large vocabulary speech recognition," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

[21] P. Karanasou, F. Yvon, T. Lavergne, and L. Lamel, "Discriminative training of a phoneme confusion model for a dynamic lexicon in asr," in *Proceedings of Interspeech*, 2013.

[22] P. Jyothi, E. Fosler-Lussier, and K. Livescu, "Discriminative training of WFST factors with application to pronunciation modeling," in *Proceedings of Interspeech*, 2013.

[23] M. Lehr and I. Shafran, "Learning a discriminative weighted finite-state transducer for speech recognition," *IEEE Transaction on Audio, Speech and Language Processing*, pp. 1360–1367, 2011.

[24] J. R. Rickford, "Phonological and grammatical features of African American Vernaculary English (AAVE)," in *African American Vernacular English*, J. R. Rickford, Ed. Malden, MA: Blackwell, 1999, pp. 3–14.

[25] W. Labov, *Language in the inner city: Studies in Black English Vernacular*. University of Pennsylvania Press, 1972.

[26] W. Wolfram, *A sociolinguistic description of Detroit Negro speech*. Arlington, VA: Center for Applied Linguistics, 1969.

[27] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen, "What kind of pronunciation variation is hard for triphones to model?" in *Proceedings of ICASSP*, 2001, pp. 577–580.

[28] B. Roark, M. Saraclar, M. Collins, and M. Johnson, "Discriminative language modeling with conditional random fields and the perceptron algorithm," in *Proceedings of the ACL*, 2007, pp. 47–54.

[29] N. F. Chen, "Characterizing phonetic transformations and fine-grained acoustic differences across dialects," Ph.D. dissertation, MIT, 2011.

[30] B. Kingsbury, H. Soltau, G. Saon, S. M. Chu, H.-K. Kuo, L. Mangu, S. V. Ravuri, N. Morgan, and A. Janin, "The IBM 2009 GALE Arabic speech transcription system," in *Proceedings of ICASSP*, 2011, pp. 4672–4675.

[31] D. Graff, J. Garofolo, J. Fiscus, W. Fisher, and D. Pallett, "1996 English Broadcast News Speech (HUB4)," Linguistic Data Consortium: LDC97S44, 1997.

[32] J. Fiscus, J. Garofolo, M. Przybocki, W. Fisher, and D. Pallett, "1997 English Broadcast News Speech (HUB4)," Linguistic Data Consortium: LDC98S71, 1998.

[33] H. Soltau, G. Saon, and B. Kingsbury, "The ibm attila speech recognition toolkit," in *IEEE Workshop on Spoken Language Technology*, 2010.

[34] L. Gillick and S. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *In International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1989, pp. 532–535.

[35] NIST, "Speech recognition scoring toolkit," http://www.itl.nist.gov/iad/mig//tools/, 2007.