PRONUNCIATION MODELING BY SHARING GAUSSIAN DENSITIES ACROSS PHONETIC MODELS

Murat Saraclar[1] *Harriet Nock*[2] *Sanjeev Khudanpur*[1]

[1] Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, USA
[2] Cambridge University Engineering Department, Cambridge, UK

ABSTRACT

Conversational speech exhibits considerable pronunciation variability, which has been shown to have a detrimental effect on the accuracy of automatic speech recognition. There have been many attempts to model pronunciation variation, including the use of decision-trees to generate alternate word pronunciations from phonemic baseforms. Use of such pronunciation models during recognition is known to improve accuracy. This paper describes the use of such pronunciation models during acoustic model training. Subtle difficulties in the straightforward use of alternatives to canonical pronunciations are first illustrated: it is shown that simply improving the accuracy of the phonetic transcription used for acoustic model training is of little benefit. Analysis of this paradox leads to a new method of accommodating nonstandard pronunciations: rather than allowing a phoneme in the canonical pronunciation to be realized as one of a few *distinct* alternate phones predicted by the pronunciation model, the HMM states of the phoneme's model are instead allowed to share Gaussian mixture components with the HMM states of the model of the alternate realization. Qualitatively, this amounts to making a soft decision about which surface-form is realized. Quantitative experiments on the Switchboard corpus show that this method improves accuracy by 1.7% (absolute).

1. INTRODUCTION

Pronunciations in spontaneous, conversational speech tend to be much more variable than in careful read speech, where pronunciations of words are more likely to adhere to their citation forms. Most speech recognition systems, however, rely on pronouncing dictionaries which contain few alternate pronunciations for most words. This failure to capture an important source of variability is potentially a significant cause for the relatively poor performance of recognition systems on large vocabulary (spontaneous) conversational speech recognition tasks. It is well known that use of a pronunciation model during recognition results in moderate improvements in word error rate (WER).

A natural extension of this idea is to incorporate the pronunciation model in the initial training of the acoustic-phonetic models. Most state-of-the-art automatic speech recognition (ASR) systems estimate these models under the assumption that words in the training corpus are pronounced in their canonical form. A word-level transcription of the speech and standard pronouncing dictionary are used to generate phone-level training transcriptions. Intuition suggests that use of a pronunciation model to improve the accuracy of this phone-level training transcription should lead to sharper acoustic models and better recognition. However, contrary to expectation and to the best of our knowledge, efforts to incorporate pronunciation modeling in acoustic model training for spontaneous speech have been unfruitful.

In this paper, we investigate this failure and consequently arrive at a novel method of pronunciation modeling. When used during recognition, our method improves accuracy to the same extent as previously used methods, and improves it even further when used in acoustic model training.

The structure of this paper is as follows. Our earlier pronunciation modeling framework is reviewed briefly in Section 2. Sections 3 and 4 investigate the straightforward approach of training acoustic models on phonetic transcriptions refined through the use of a pronunciation model. This leads to little improvement in WER. Section 5 considers direct use of hand-labeled transcriptions to further bootstrap the acoustic model training process of Section 3. In an apparent paradox, the acoustic models resulting from these procedures degrade WER but the phone-accuracy of the resulting word-hypotheses is actually better than that of the baseline system. This leads, in Sections 6 and 7, to a new way of capturing pronunciation variation, dubbed *state-level pronunciation modeling* (as opposed to the preceding *phone-level* pronunciation model).

2. PRONUNCIATION MODELING FRAMEWORK

We begin with a very brief review of our pronunciation modeling methodology (see [1] for details). The main steps in using a pronunciation model for ASR are to

1. **obtain a canonical (phonemic) transcription** of some training material. A standard pronouncing dictionary (in our case, PronLex) is used for this purpose.
2. **obtain a surface-form (phonetic) transcription** of the same material. A portion of the Switchboard corpus has been phonetically hand labeled by linguists (see [4]).
3. **align the phonemic and phonetic transcriptions**. A dynamic programming procedure based on phonetic feature distances is used for this purpose.
4. **estimate a decision-tree pronunciation model**. A decision tree is constructed to predict the surface form of each phoneme by asking questions about its phonemic context.
5. **perform recognition with this pronunciation model**. The pronunciation model is used to transform each phoneme in a dictionary-based phoneme-level recognition network to yield a network of surface-forms. Recognition is performed on this phone-level network. The phoneme-level network may either be derived from a word-level language model or from a word-lattice generated by an initial recognition pass.

It is shown in [3] that if only a small amount of phonetically labeled data is available in Step 2, the pronunciation model in Step 4 and the corresponding WER in Step 5 *are worse* (1.4% absolute) than using canonical pronunciations. One way to automatically generate more data for Step 2 is

6. **full training set retranscription**. Starting with the canonical transcription of the entire acoustic training set (instead of just the hand-labeled portion in Steps 1-2), the pronunciation model of Step 4 is used to create pronunciation networks representing possible phonetic realizations of each training utterance. The most likely phone-sequence through each network is chosen via Viterbi alignment using a set of *existing* acoustic models, giving a "refined" transcription of the entire training set.

[0]This work was partially supported by the National Science Foundation under Grant No. 9714169

It is shown in [3] that replacing the small corpus of Step 2 with the larger corpus of Step 6, and then repeating Steps 3-5 leads to a small but statistically significant ($\sim$ 0.5% absolute) improvement in WER on the Switchboard corpus.

3. IMPROVING THE PHONETIC TRANSCRIPTIONS USED IN ACOUSTIC TRAINING

The "refined" transcriptions resulting from Step 6, it may be reasoned[1], are better suited for acoustic model training than the canonical baseforms. This leads to a notion of

7. **acoustic model reestimation**. *New* acoustic models are reestimated based on the phone transcriptions of Step 6.

The retranscription of Step 6 is then repeated with these *new* acoustic models replacing the *existing* acoustic models used earlier. The resulting phonetic transcription[2] is then used in Steps 3-5 for pronunciation model estimation and recognition.

It is possible to gauge the quality of the phonetic transcription of Step 6 by comparing with hand labels which are available for a portion of our Switchboard training corpus. Table 1 presents this comparison for 1800 sentences (40,000 phones). It is clear from

Transcriptions	Phone Error Rate
Dictionary Baseforms	28.3%
Automatic (Step 6)	26.1%

Table 1. Training Transcriptions as compared to Hand-Labels

Table 1 that the models in Step 7 are trained on more accurate phonetic transcriptions. However, they result in exactly the same recognition performance (38.9% WER) as the acoustic models trained on canonical baseforms!

The first hypothesis we investigate is that although the transcriptions resulting from Step 6 are closer to the hand-labels, they still contain many inappropriate phones due to poor phone recognition performance in Step 6.

Method We attempt to improve the quality of phone recognition by standard speaker adaptation techniques. Vocal Tract Length Normalization (VTLN) and Maximum Likelihood Linear Regression (MLLR) are used to adjust the acoustic models before performing the retranscription in Step 6.

Adaptation Method	Phone Error Rate
ML-VTLN	26.0%
MLLR	26.0%

Table 2. Failed Attempts to Improve Training Transcriptions

Results The use of adaptation techniques leads to little change in transcription accuracy relative to the hand-labeled transcriptions (Table 2). It also results in little change in transcription content as evidenced by the comparison of the three automatic transcription techniques in Table 3. The new transcriptions remain fairly close to the original baseform transcriptions both before and after adaptation.

Discussion The results suggest the original hypothesis – that the Step 6 transcription was poor due to low phone-recognition accuracy – is incorrect; we conclude instead that the highly-parameterized set of acoustic models used here is so well-tuned to the PronLex baseforms on which it is trained that little change in the transcriptions can be obtained when using these models for the retranscription stage. Adaptation based on the training transcriptions simply exacerbates the problem. This conclusion

Acoustic Models Used in Step 6	Phone Error Rate Relative to	
	Prev Automatic Transcriptions	Baseforms
Baseline (unadapted)	0.0%	4.1%
ML-VTLN adapted	0.7%	4.2%
MLLR adapted	1.5%	4.0%

Table 3. Automatic Transcriptions Before/After Adaptation

Acoustic Model Used in Step 6	Phone Error Rate
8-Gaussian triphone models	25.7%
1-Gaussian triphone models	25.5%

Table 4. Simpler Acoustic Models Improve Transcription

is supported by a small increase in phone transcription accuracy with respect to the hand-labeled data when retranscription uses acoustic models of lesser complexity (Table 4).

4. JACK-KNIFING TO FURTHER IMPROVE THE ACOUSTIC TRAINING TRANSCRIPTIONS

Since transcription accuracy is improved when retranscription uses smaller models trained on the same data set, a natural progression is to retranscribe the training set using models trained on different data.

Method The 60-hour Switchboard training set is partitioned into two speaker disjoint gender-balanced 30 hour subsets and model sets trained on one half phonetically transcribe the acoustics for the unseen half of the data (as in Step 6). The resulting transcriptions are then used to train a set of acoustic models (as in Step 7). Steps 6, 3, 4 and 5 are then carried out to estimate and test a pronunciation model.

Results The phone recognition accuracy relative to the hand-labeled transcriptions is essentially unchanged by the cross-transcription method (25.3%). This is not to say that the resulting transcriptions are the same as those described in the preceding section. Indeed these transcriptions deviate even more from the baseforms than the transcriptions of Table 3. Despite this, the "refined" transcriptions do not lead to any significant change in recognition performance (38.9% WER).

Discussion We conclude that it is quite difficult to obtain accurate *automatic* phonetic transcriptions using acoustic models which are trained on canonical baseforms.

5. USING ACOUSTIC MODELS TRAINED ON HAND-LABELED DATA

One way to obtain more accurate phonetic transcriptions of the entire acoustic training corpus (Step 6) is to use acoustic models which are trained directly on only the hand-labeled portion of the training corpus. We investigate this avenue as well.

Method Only a small portion (3.5 hours) of the acoustic training data has been transcribed at the phone level by human labelers. Due to this limitation, we estimate a set of context-independent phone models (called *ICSI-models*) using the hand-labeled portion of the training set.

Step 6 is performed next, replacing the *existing* acoustic models with the *ICSI-models*. This results in considerably more accurate phonetic training transcription (see results below). Step 7, training acoustic models on the entire training set, is performed next. The resulting models are named *ICSI-bootstrap models*. This is followed by the usual procedure (Steps 6, 3, 4, and 5) of estimating and testing a new pronunciation model appropriate for these acoustic models.

[1] The "refined" phonetic transcriptions agree better with the hand-labels than the canonical baseforms, and are therefore more "accurate" for training acoustic phonetic models.

[2] Redoing steps 6, 3-5 after Step 7 ensures that the final pronunciation model in Step 4 is matched to the *new* acoustic models used in recognition in Step 5.

Results First we present results showing that phone transcription accuracy is improved by models trained on hand labels. Since these models are bootstrapped from the phonetically labeled training utterances on which the results of Tables 1-4 are reported, it is inappropriate to compare transcription accuracy on that set. We therefore use a 451-utterance subset of our test set, which also has phonetic labels, to compare the transcription accuracy of the *ICSI-models* with models trained on canonical pronunciations. The task is the same as Step 6: choose the best phone-sequence given the word transcription and a pronunciation model. The results of Table 5 for the *ICSI-models* indicate that the transcriptions on which the *ICSI-bootstrap* models are trained are likely to be much more accurate than the baseforms or the transcriptions used in preceding sections. The *ICSI-bootstrap* models also *appear* to be considerably better phonetic-models than standard models trained on canonical baseforms.

Transcription Type	Models	Phone Error Rate
Dictionary Baseforms	—	33.6%
Automatic (Step 6)	Standard	31.4%
Automatic (Step 6)	ICSI-models	26.6%
Automatic (Step 6)	ICSI-bootstrap	26.6%

Table 5. Using Hand Labeled Data to Train Acoustic Models for Improved Phone Transcription given the Word Transcription

The recognition performance however turns out quite the contrary. While the standard acoustic models (together with a pronunciation model) have a WER of 38.9%, the WER of the *ICSI-bootstrap* models is 41.3%! In order to better understand the cause of this degradation, the performance of the model on the 451 phonetically labeled utterances in the test data is analyzed. In addition to the WER performance the phone error rate is measured against the hand transcriptions. It turns out (Table 6) that the *ICSI-bootstrap* models improve phone accuracy by 4.5% on this subset of the test set, although the WER is worse by 1.4%.

Pronunciation Model Used in Step 5 (Test)	Acoustic Model			
	Standard		ICSI-bootstrap	
	PER	WER	PER	WER
None (Dictionary)	49.1%	49.1%	49.5%	58.9%
Tree Pron. Model	47.7%	48.7%	43.2%	50.1%

Table 6. Comparison of Word and Phone Error Rates for Different Acoustic and Pronunciation Models

Discussion It is clear from these experiments that there indeed is considerable deviation from canonical pronunciations in spontaneous speech and that the *ICSI-bootstrap* models are indeed better at capturing the actual realized pronunciations than models trained on standard pronunciations. We conclude that the inability to translate this (implicit) lower phone error rate into lower WER is due to lexical confusion: since our decision-tree pronunciation model allows words to have a large number of pronunciations, many of which overlap with pronunciations of other words, "recovering" the right word strings from the more accurate phone recognition is difficult. Yet, the model for the acoustic realization of a phoneme must allow for the inherent variability. This leads to a new way of modeling pronunciations.

6. MODELING PRONUNCIATION VARIABILITY AT THE LEVEL OF HMM STATES

In this section we present a new way to model alternate pronunciations and show that it performs as well as the decision tree pronunciation model described in [3]. This new model accommodates alternate surface-form realizations of a phoneme by allowing the HMM state of the model of the phoneme to share output densities with models of the alternate realizations. We call this a *state level pronunciation model* (SLPM) for reasons described below.

To understand the SLPM, first consider the effect of a more traditional pronunciation model which allows the baseform /abc/ to be alternately realized as the surface form /asc/. The sketch at the top of Figure 1 illustrates how a context-independent HMM system will permit this alternative in the recognition network, and the sketch in the middle illustrates the same for a context-dependent (triphone) HMM system. The SLPM deviates from these methods as illustrated by the sketch at the bottom of Figure 1. Rather than letting a phoneme /b/ be realized as an alternate phone /s/, the HMM states of the acoustic model of the phoneme /b/ are instead allowed to utilize the output density of the HMM states of the acoustic model of the alternate realization /s/. Thus the acoustic model of a phoneme /b/ has the canonical and alternate realizations (/b/ and /s/) represented by different sets of mixture components in one set of HMM states.

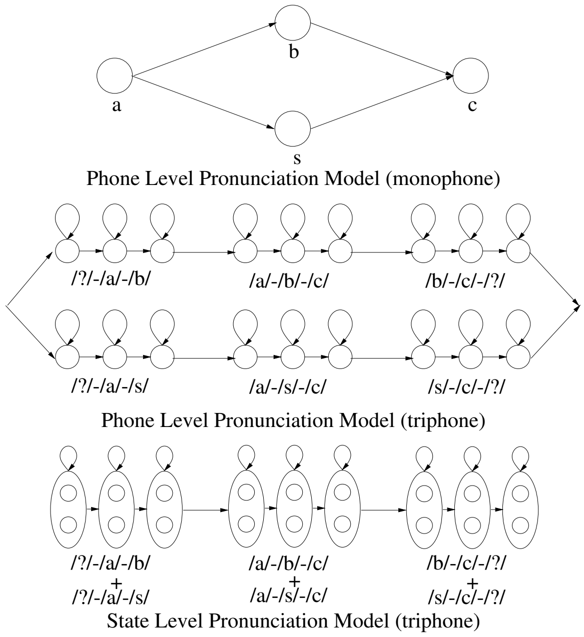

Figure 1. The Effect of Allowing a Phoneme /b/ to be Realized as a Phone /s/, Viewed at the Level of HMM States

Method To construct an HMM system in which states share output densities based on a pronunciation model we

1. Obtain a state to state alignment between the baseform and surface form representations (similar to Step 3, Section 2).
2. Estimate the probability of a HMM *state* being realized as an alternate *state*, using this alignment, Prob$(/\mathtt{s}/|/\mathtt{b}/)$.
3. Filter out unreliable estimates.
4. Modify the output distribution of the baseform state to include the mixture components of the alternate states.
5. Further train the resulting "tied-mixture"-like acoustic models.

The state output densities in our system are mixtures of Gaussians:

$$P(o|/\mathtt{b}/) = \sum_{i \in G(/\mathtt{b}/)} w_{i,/\mathtt{b}/} \mathcal{N}(o; \mu_i, \Sigma_i)$$

where $G(/\mathtt{b}/)$ denotes the set of mixture components $\mathcal{N}(o; \mu_i, \Sigma_i)$ for state $/\mathtt{b}/$ and w_i denotes their mixture weights. In the example of Figure 1, Step 4 in the SLPM construction replaces G with G' and w with w' where

$$\begin{aligned} G'(/\mathtt{b}/) &= G(/\mathtt{b}/) \cup G(/\mathtt{s}/) \\ w'_{i,/\mathtt{b}/} &= \text{Prob}(/\mathtt{b}/|/\mathtt{b}/) w_{i,/\mathtt{b}/} + \text{Prob}(/\mathtt{s}/|/\mathtt{b}/) w_{i,/\mathtt{s}/}. \end{aligned}$$

This formalism extends easily to the more general case of a phoneme that may be realized as one of several alternate phones.

Results The SLPM developed above is used in a recognition experiment on the Switchboard corpus[3]. Table 7 shows that just as the decision tree pronunciation model, the SLPM results in a small but significant reduction in WER.

Pronunciation Model	WER
None (PronLex Dictionary)	39.4%
Decision Tree Pronunciation Model	38.9%
State Level Pronunciation Model	38.8%

Table 7. Recognition Performance of the SLPM

Discussion It may be noted that the Gaussian densities are not duplicated before reestimation, but are shared among states. The only additional parameters introduced are the mixture weights. This increases the number of free parameters in the acoustic models by less than 0.5%.

The SLPM mimics the behavior of our decision tree pronunciation model with some additional advantages.

- Canonical (phonemic) transcriptions can be used to train HMMs resulting from the SLPM construction. Since output densities of the alternate realizations are present in the HMM state of the canonical pronunciation, acoustic realizations which match the alternate phones better will be used by the Baum-Welch reestimation to update those densities instead of the canonical ones.
- The dictionary need not be expanded to include alternate pronunciations, an important consideration for recognition speed.

7. INTRODUCING MORE ACCURATE DENSITIES TO MODEL THE SURFACE-FORM REALIZATIONS

The essential idea of the SLPM in the previous section is to augment the HMM state of a phoneme so as to model alternate surface-form realizations. The origin of the densities which model the alternate realizations merits further discussion. In Figure 1, when HMM states of `/b/` needs to account for a realization `/s/`, densities of an HMM for `/s/` *trained on canonical transcriptions* is used. An alternative is to instead augment the HMM states of `/b/` with densities from HMM for `[s]` *trained on the more accurate transcriptions* of Section 5.

Method The only modification to the SLPM recipe of Section 6 is that in Step 4, the output densities are augmented by sharing mixture components not only with existing HMMs, but also [4] with densities from the corresponding HMM states of Section 5. This is illustrated in Figure 2. Further training of the models (cf. Step 5 in Section 6) is achieved by first training the mixture weights and transition probabilities followed by training the whole model.

Results The results in Table 8 indicate that this modified HMM set performs significantly better than HMMs trained on canonical pronunciations, giving a gain of 1.7% (absolute) in WER. When two sets of acoustic models are "merged" in this fashion, the number of parameters is nearly doubled. One way to make a fair comparison is to compare the "merged" SLPM system with a system that has 24 Gaussians per state. However, data sparseness causes the 24 Gaussians-per-state system to be over trained and its WER on the test set is 39.7% which is even worse than the 12 Gaussians per state baseline.

[3]The baseline acoustic models are state-clustered cross-word triphone HMMs [5] having about 6700 states each with 12 Gaussian densities per state. The language model is a trigram trained on about 2.2 million words, and the front-end uses MF-PLP derived coefficients. The test set has 19 conversations, amounting to about 2 hours of speech with about 18000 words.

[4]Note that this does substantially increase the number of parameters in the system.

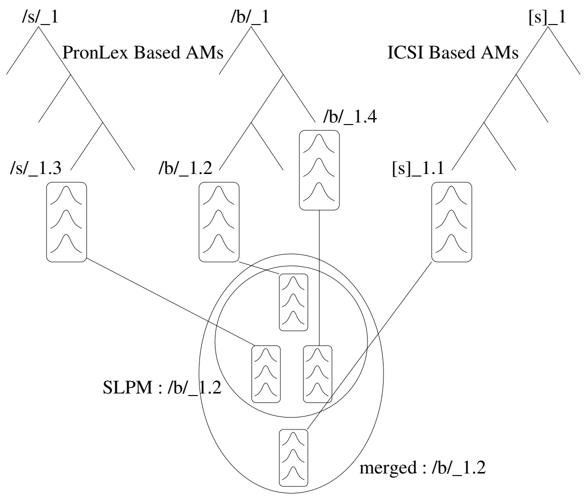

Figure 2. Merging Gaussian Mixtures

Pronunciation Model	Acoustic Model	WER
PronLex (baseline)	PronLex based	39.4%
Phone Level PM	PronLex based	38.9%
State Level PM	merged, no training	38.2%
State Level PM	merged, further training	37.7%

Table 8. Performance of acoustic model merging

Discussion In another effort to make a fair comparison by keeping the number of parameters in the SLPM comparable to the baseline models, two sets of acoustic models with a smaller number of mixture components (6 per state) are merged. The resulting retrained system has a WER of 38.3%, which is still substantially better than the decision tree pronunciation model.

8. ACKNOWLEDGEMENTS

The authors thank Michael Riley at AT&T Laboratories, for providing the FSM tools required to manipulate the pronunciation models and Entropic Cambridge Research Laboratory, UK, for providing the lattice generation tools used in these experiments. The idea of sharing output densities was inspired by the doctoral dissertation of Xiaoqiang Luo [6].

REFERENCES

[1] M. Riley and A. Ljolje, "Automatic generation of detailed pronunciation lexicons." *Automatic Speech and Speaker Recognition: Advanced Topics*. Kluwer. 1995.

[2] W. Byrne, *et al*, "Pronunciation Modelling for Conversational Speech Recognition: A Status Report from WS97," presented at the 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara, CA, Dec. 1997.

[3] W. Byrne, *et al*, "Pronunciation Modelling Using a Hand-labelled Corpus for Conversational Speech Recognition", in *Proc. ICASSP '98* Seattle, WA, May 1998.

[4] S. Greenberg, "The Switchboard Transcription Project," *1996 LVCSR Summer Workshop Technical Reports*, 1996, `http://www.icsi.berkeley.edu/real/stp/`

[5] S. Young, J. Jansen, J. Odell, D. Ollasen, P. Woodland, *The HTK Book (Version 2.0)*, Entropic Cambridge Research Laboratory, 1995.

[6] X. Luo, "Balancing Model Resolution and Generalizability in Large Vocabulary Continuous Speech Recognition." PhD Thesis, The Johns Hopkins University, Baltimore, MD, 1999.