Computer Speech and Language (2000) **14**, 137–160
Article No. 10.1006/csla.2000.0140
Available online at http://www.idealibrary.com on IDEAL

Pronunciation modeling by sharing Gaussian densities across phonetic models

Murat Saraçlar,†§ Harriet Nock‡¶ and Sanjeev Khudanpur†§

†Center for Language and Speech Processing, The Johns Hopkins University, Baltimore, MD, U.S.A., ‡Cambridge University Engineering Department, Trumpington Street, Cambridge, U.K.

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 pronunciation models during recognition is known to improve accuracy. This paper describes the incorporation of pronunciation models into acoustic model training in addition to recognition. 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. Acoustic models trained on the most accurate phonetic transcriptions result in worse recognition than acoustic models trained on canonical baseforms. Analysis of this counterintuitive result 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, the hidden Markov model (HMM) states of the phoneme's model are instead allowed to share Gaussian mixture components with the HMM states of the model(s) of the alternate realization(s). Qualitatively, this amounts to making a soft decision about which surface form is realized. Quantitatively, experiments show that this method is particularly well suited for acoustic model training for spontaneous speech: a 1.7% (absolute) improvement in recognition accuracy on the Switchboard corpus is presented.

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, both for training and recognition. 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)

§E-mail: {murat,sanjeev}@cslp.jhu.edu
¶E-mail: {hjn11}@eng.com.ac.uk

0885–2308/00/020137 + 24 $35.00/0

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). There have been fewer attempts 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 acoustic models under the assumption that words in the training corpus are pronounced in their canonical form. A word-level transcription of the speech is used along with a standard pronouncing dictionary 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 only during recognition and not in acoustic model training, our method improves accuracy to the same extent as previously used methods for pronunciation modeling. When used for acoustic model training, it improves accuracy even further.

The structure of this paper is as follows. After a brief review of the corpus, task and baseline system in Section 2, empirical evidence is presented in Section 3 to motivate the need for pronunciation modeling for spontaneous speech in general and for this task in particular. A short summary of directly related work by other researchers follows in Section 4 to provide the backdrop for this research. A specific pronunciation modeling methodology whose details are relevant to this paper is then reviewed briefly in Section 5. The two main contributions of this paper appear in Sections 6 and 7.

In Section 6, we investigate several ways to improve the accuracy of the phonetic transcriptions used for training acoustic models for spontaneous speech. We find that small improvements in training transcription accuracy lead to little improvement in the recognition WER of the resulting system. In an apparent paradox, we find that a more significant improvement in the accuracy of the phonetic transcriptions used for acoustic training *degrades* recognition WER relative to the baseline system! Yet, the phone accuracy of the resulting recognizer output is actually higher than that of the baseline system. Further analysis of this result motivates the need for a method of modeling pronunciation variations in which *no additional homophony is introduced by adding new pronunciations to the recognizer's dictionary and yet alternate acoustic realizations are somehow permitted*.

In Section 7, we present such a method of capturing pronunciation variation in which HMM states of phonemes share output densities with HMM states of the alternate realizations. We call this novel method *state-level pronunciation modeling* to contrast it with more traditional *phone-level* pronunciation models.

Studying the results of Section 6 is recommended for a deeper understanding of Section 7. However, the introductory part of Section 6 (preceding Subsection 6.1) may be adequate for a reader who wishes to proceed directly to Section 7.

2. Corpora, task and system definition

Corpus and task description. Our main interest is modeling the pronunciation variations in large vocabulary conversational speech. Switchboard, a corpus of spontaneous telephone conversations between two individuals about loosely specified topics such as AIDS, gardening, or health-care for the elderly, is used in our experiments. [See Godfrey, Holliman and McDaniel (1992) for a detailed corpus description.] A vocabulary of approximately 20 000

words provides adequate coverage for the corpus. We use 60 hours of speech (about 100 000 utterances or a million words) selected from about 2000 conversations for acoustic model training purposes. There are 383 different speakers in the training corpus. A speaker-disjoint set of about 1.5 hours of speech (19 entire conversations, 2427 utterances, 18 100 words) is set aside for testing ASR systems.

Manual phonetic transcriptions produced by Greenberg (1996) at ICSI are available for a portion of this speech corpus. This hand labeled portion, henceforth referred to as the ICSI transcriptions, includes a 3.5 hour subset (3600 utterances, 100 000 phones) of the training set and a 0.5 hour subset (451 utterances, 18 000 phones) of the test set.

Finally, another speaker-disjoint test set of about half an hour of speech (6 conversations, 882 utterances, 6350 words) is used for a one-time blind evaluation at the end of the paper: except for computing the WER, no other diagnostics are performed on this set.

Baseline ASR system and recognition experiments. Our baseline acoustic models are state-clustered cross-word triphone HMMs having 6700 shared states, each with 12 Gaussian densities per state. The PronLex dictionary (PronLex, 1995), which has a single pronunciation for approximately 94% of the words in the test vocabulary, two pronunciations for more than 5% of the words and three or four pronunciations for the remaining (less than 0.5%) words, is used in the baseline system. Bigram and trigram models trained on 2.2 million words of transcribed Switchboard conversations are used as language models.

For speech recognition experiments we first generate word lattices using the baseline system with a bigram language model. These lattices are then used as a word graph to constrain a second recognition pass in which a trigram language model is used. We chose to use this lattice-rescoring paradigm for fast experimentation while allowing a search over a large set of likely word sequences.

Acoustic model training and lattice rescoring is carried out using the HTK HMM toolkit developed by Young, Jansen, Odell, Ollasen and Woodland (1995). The AT&T Weighted Finite State Transducer tools provided by Mohri, Pereira and Riley (2000) are used to manipulate word and phone lattices.

Performance measures. The most common measure of performance for an ASR system is WER. State-of-the-art ASR systems achieve 30–35% WER on the Switchboard corpus. The baseline system described above has comparable performance. The best possible WER obtainable from hypotheses in the word lattices we use is less than 10%. This is adequate for experiments within the lattice-rescoring paradigm.

In addition to WER, we use phone error rate (PER) as another measure of transcription accuracy. PER will be reported later in this paper on the 451 test set utterances (or 1800 of the 3600 training set utterances) for which the correct phonetic transcription is available.

3. Motivation for pronunciation modeling

Conversational or casual speech differs from formal or read speech in many aspects, a very remarkable one being the deviation from canonical pronunciation of words. Some causes of pronunciation variation such as a speaker's accent or dialect are common to conversational and read speech. Others are characteristic of casual speech. For example, some deviations may be attributable to coarticulation of words, e.g. `hafto` instead of `have to`. Others are due to commonly acceptable reductions or due to the fast nature of conversational speech, e.g. `wanna` instead of `want to` or `dijha` instead of `did you`.

A study by Bernstein, Baldwin, Cohen, Murveit and Weintraub (1986) reveals that typical conversational speech, to begin with, is faster than typical read speech in terms of the average number of words spoken per unit time. However, if one measures the speaking rate in terms of the number of *phones* per unit time, then spontaneous speech is very comparable to read speech! This suggests that pronunciations in spontaneous speech are different from read speech, in that speakers tend to delete phones rather than merely reduce phone durations during spontaneous speech. Furthermore, deletion of a phone is often accompanied by suitable modification in adjacent phones so as to preserve intelligibility. In Fosler *et al.* (1996), we have compared the ICSI phonetic transcriptions of a portion of the Switchboard corpus with the citation-form pronunciations of the transcribed words and found that 12.5% of the phones in the standard (PronLex) pronunciation are deleted; substitutions and insertions of phones also change pronunciations so that only 67% of the phones in the "prescribed" pronunciations are correctly articulated by an average speaker. This has a significant impact on the performance of ASR systems which are typically built on the assumption that speakers only use canonical pronunciations.

3.1. Recognizer degradation with differing speaking style

We performed an extended version of an experiment reported earlier by Weintraub, Taussig, Hunicke-Smith and Snodgrass (1996) to investigate whether casual speaking style *is indeed* a major factor contributing to the poor recognizer performance on spontaneous speech tasks. Our experiment used the MULTI-REG corpus collected by Weintraub *et al.* (1996), which comprises conversations recorded in different speaking styles. About 15 *spontaneous* telephone conversations between individuals were collected first. These telephone-bandwidth (narrow-band) recordings were similar to those made when compiling the Switchboard corpus. A simultaneous recording of the speech using wide-band close-talking microphones was also made. The subjects were later recalled to make two further recordings based on transcripts of their original spontaneous conversations: first *reading* a transcript as if dictating it to a computer and then re-reading the same transcript *imitating* a conversation. Simultaneous telephone quality *narrow-band* and high-quality *wide-band* stereo recordings were made for these two speaking styles as well. These six renditions of the same conversation, controlled for two principle axes of variability, speaking style and recording bandwidth, were used in our experiment.

Two large-scale HTK-based recognizers, one trained on narrow-band Switchboard data and the other trained on wide-band Wall Street Journal (WSJ) data, were used to recognize the MULTI-REG spoken utterances which have identical word-level reference transcriptions across the six conditions. The results are shown in Table I. The first column reconfirms the result of Weintraub *et al.* (1996) that narrow-band models trained on spontaneous speech perform much better on read and imitated-spontaneous speech than on truly spontaneous speech, despite the mismatch between the speaking styles in training and test data. The table also shows the same trend when the experiment is repeated with wide-band models and data. (Note that although *degradation* in results is comparable between wide- and narrow-band experiments, the error rate results are not directly comparable due to differences in model parameterization.)

The decrease in accuracy with increasingly casual speaking style is seen across both bandwidths, whilst the handset and words pronounced in the test data remained unchanged. It therefore seems likely that the degradation is due to changes in speaking style, and the table further suggests that training and testing on data with matched speaking styles offers only

TABLE I. WER degradation with speaking style on the MULTI-REG test set

Speaking style	Narrow-band (SWBD models)	Wide-band (WSJ models)
Reading	26.1%	26.2%
Imitating	29.5%	39.7%
Spontaneous	43.2%	62.4%

partial robustness to style changes. One factor contributing to the degradation is likely to be the increased variability in pronunciation in fluent speech, providing further motivation for modeling pronunciation variation.

As an aside, note that the error rate for read speech under WSJ models in Table I is higher than those typically reported on WSJ test sets. The transcripts of spontaneous conversations being read here, however, contain many short words and short words have been shown to be more error-prone (Eide, Gish, Jeanrenaud & Mielke, 1995).

3.2. Cheating experiments

In order to gauge the maximum improvement in recognition accuracy achievable if one could predict the actual pronunciations that were used by a speaker in the test set, we conducted some "cheating" experiments. We fixed a set of acoustic phonetic models and performed unconstrained phone recognition on the *test speech*. We then aligned the phone string resulting from this process with the reference word transcriptions for the test set (hence the cheating) and extracted the *observed* pronunciation of each word in the test set. Many of these pronunciations were different from the canonical pronunciations. Note that since automatic means were used for phone transcription, the resulting pronunciations of the words are not the same as those one would infer from a manual phonetic transcription of the same speech.

With these alternative pronunciations at hand, we enhanced the pronunciation dictionary used during recognition. This procedure had two variants:

- New pronunciations of all words encountered in the entire test set were added to a static pronunciation dictionary. Except for some coarticulation effects, the best a pronunciation model can do is predict those and only those new pronunciations of a word which are actually seen in the test data.
- The pronunciation dictionary was modified individually for each test utterance. This is almost *the best* a pronunciation model can do, including coarticulation effects, because a majority of words are seen only once in an utterance.

These dictionary enhancements were used to rescore lattices obtained using an ASR system[1] with a WER of 47%. The static dictionary enhancement reduced the WER to 38% and the utterance based enhancement of the dictionary to 27%. The paths with the least error rate in the lattices we used in this experiment (determined by looking at the correct transcription) had 13% WER.

These experiments therefore provide a very high margin for improvement which is possible, or at least not ruled out, if one can accurately predict word pronunciations. In addition,

[1]This ASR system, developed when we started our research, is older than the baseline system used through out the remainder of the paper and its WER is significantly worse. The relative change in WER in this experiment motivates pronunciation research. However, comparison of the absolute WER of this ASR system and that of the baseline system should not be made.

the enhanced dictionaries used in these experiments did not have probabilities assigned to each entry. The actual pronunciation models we use do assign probabilities to each alternate pronunciation. Thus, in fact, it is possible to obtain even lower WERs than these "lower bounds."

Similar experiments were conducted by McAllaster, Gillick, Scattone and Newman (1998) using simulated data. Their results also support the hypothesis that knowing the correct pronunciations can result in large gains.

4. Related work

Elaborate pronunciation modeling for conversational speech recognition has received attention only in the last few years. Much of the work prior to that has been on read speech tasks and in the spirit of incremental changes to the permissible pronunciations of select individual words. We briefly review some work on pronunciation modeling for large vocabulary spontaneous speech, specifically focusing on North American English. The reader is referred to Strik and Cucchiarini (1999) for a more comprehensive review of other work in this field.

Modeling alternate pronunciations of entire words has been attempted on spontaneous speech tasks. Peskin *et al.* (1999) and Hain and Woodland (1999*a*) use pronunciation probabilities in their baseform dictionary, which may be considered a weak form of pronunciation modeling. Sloboda and Waibel (1996) generate alternate pronunciations using an automatic phone recognizer on a training corpus to obtain frequent alternative pronunciations of frequent words. This approach, besides relying on having a good phone recognizer, cannot predict unseen pronunciations or pronunciations for words not seen in the acoustic training set (often referred to as *unseen words*).

An alternative to modeling pronunciation variation of entire words is to view the new pronunciations as being obtained by local changes in the phonemes of a standard pronunciation and to model the process of this change, e.g. the pronunciation `[h ae f]` of the word `have` is obtained by changing the third phoneme in its baseform `[h ae v]`. These approaches can generate pronunciations for all words including the unseen words. Techniques which espouse this idea may be divided into those which start with hand-crafted rules and those which infer the rules of pronunciation change from data.

Cohen (1989), Giachin, Rosenberg and Lee (1990) and Tajchman, Fosler and Jurafsky (1995) have used phonological rules obtained from linguistic studies to generate alternative pronunciations for read speech tasks. Finke and Waibel (1997) applied these methods to spontaneous speech, and further extended them by making the pronunciation probabilities depend on dynamic features such as speaking rate, segment durations and pitch. The main limitation of these procedures is the need for hand-crafted rules, and the consequent inability to model observed changes which are not covered by a rule in the inventory.

Lucassen and Mercer (1984), Chen (1990) and Riley (1991) have used statistical decision trees to generate alternate word pronunciations for read speech tasks. In Fosler *et al.* (1996) and Byrne *et al.* (1997), we extended this work to spontaneous speech. Fosler-Lussier (1999) also used these techniques to predict alternate pronunciations of syllables and words instead of phonemes.

Very recently, Finke, Fritsch, Koll and Waibel (1999) proposed a new technique that attempts to model a phoneme as a feature bundle which can be augmented due to a pronunciation change. In contrast to all the techniques mentioned earlier, which assume a *complete* change in the identity of a phonetic segment when a pronunciation change occurs, this

technique is able to model partial changes. The method presented in this paper shares this characteristic of modeling partial pronunciation changes.

A technique proposed by Wooters and Stolcke (1994) for generating a phone graph for a word from empirically observed pronunciations has been extended by Eide (1999) to generating a HMM state graph for a context-dependent phoneme. Wakita, Singer and Sagisaka (1999) also use empirically observed HMM state sequences to infer alternate word pronunciations at the granularity of HMM states instead of phonemes. The method presented in this paper shares this characteristic of these two methods: modeling pronunciation change at the level of a HMM state.

All the pronunciation models mentioned above result in a small reduction in WER, compared to using a baseform dictionary, when used during recognition. However, very few attempts to use the pronunciation model in conjunction with acoustic model training have been reported. Sloboda and Waibel (1996) claim that the use of alternate pronunciations for acoustic model training provides improvement beyond the gain from using them during recognition alone, but their experimental results are inconclusive. Finke and Waibel (1997) use alternate pronunciations during acoustic model training, but this is done in conjunction with "flexible transcription alignment." As a consequence, it is difficult to estimate the part of the overall gain which may be attributed to the use of pronunciation modeling in acoustic model training.

5. Pronunciation modeling framework

We begin with a review of the decision-tree-based pronunciation modeling methodology originally developed by Riley and Ljolje (1995). We essentially adapt the techniques they developed for read speech to our spontaneous speech task. This brief review is included for the sake of completeness and interested readers are referred to Riley *et al.* (1999) for details.

The main steps in training and using a pronunciation model for ASR, as shown in the block diagrams of Figures 1, 2 and the example of Figure 3, 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, with Viterbi alignment when there are multiple pronunciations.
(2) **obtain a surface-form (phonetic) transcription** of the same material. The portion of the Switchboard corpus that is phonetically hand labeled by linguists is used (see Greenberg, 1996).
(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**. A dictionary-based phoneme-level recognition network is first obtained from a word lattice generated by an initial recognition pass. The pronunciation model is then used to transduce phoneme sequences in this network to yield a network of surface-form realizations. Figure 4 illustrates this expansion for a two-word fragment. Recognition is then performed on this phone-level network using the baseline acoustic model trained from the phonemic transcripts.
(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 represent-

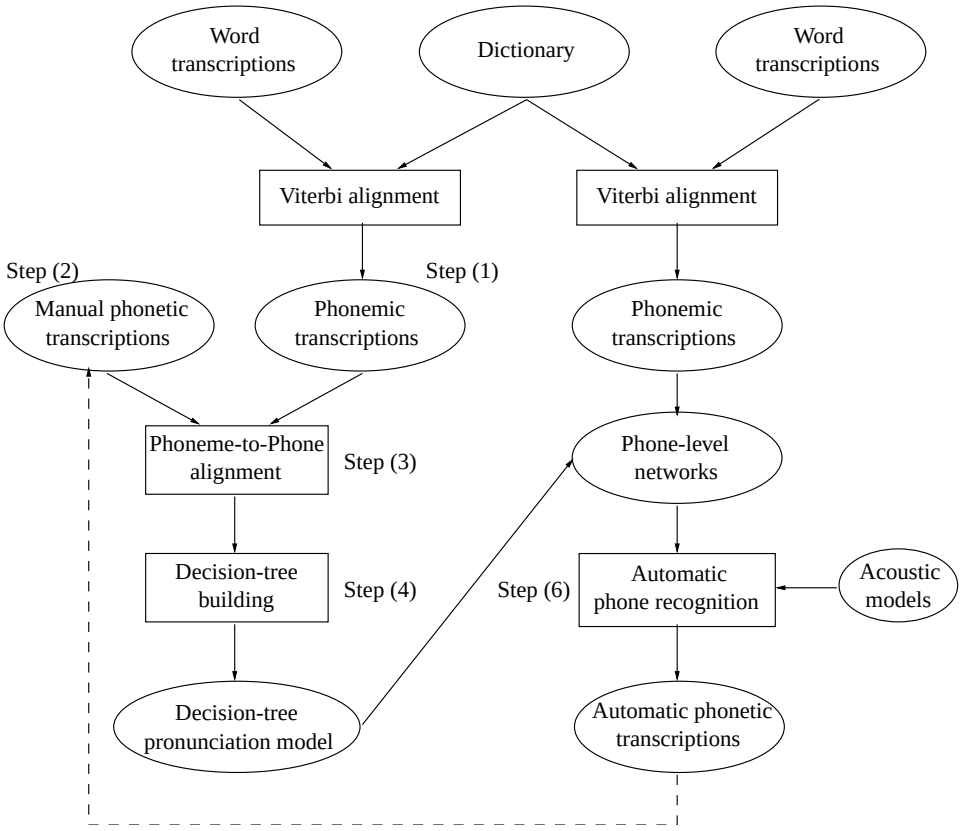

Figure 1. Pronunciation model training.

ing 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.

We have shown in Byrne *et al.* (1998) and Riley *et al.* (1999) 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. We conjecture that the mismatch between the acoustic models and the hand transcriptions as well as the high degree of lexical confusion are the main reasons for this degradation. One way to automatically generate more data for Step (2) is to replace the small corpus of Step (2) with the larger corpus of Step (6), and then repeat Steps (3)–(5). This leads to a small but statistically significant (39.4% $\rightarrow$ 38.9%, $\sim$0.5% absolute) improvement in WER on the Switchboard corpus.

Note that in the Switchboard results reported in both Byrne *et al.* (1998) and Riley *et al.* (1999), canonical pronunciations of words were used for estimating acoustic phonetic models. The automatic phonetic transcription of the acoustic training corpus generated in Step (6) was used only to estimate a pronunciation model and alternate pronunciations generated by this model were permitted only during recognition. We show next, in Section 6, that the phonetic transcription of Step (6) is more accurate than a transcription based on canonical

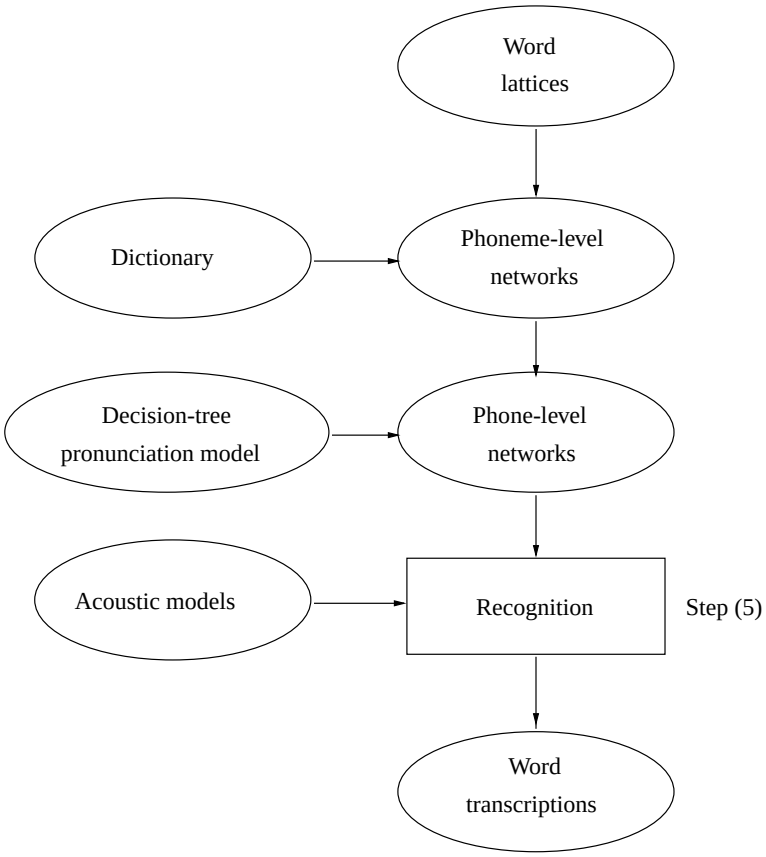

Figure 2. Recognition using the pronunciation model.

TABLE II. Improved training transcriptions for acoustic model estimation

Transcriptions	Phone error rate	Deviation from baseforms
Dictionary baseforms	28.3%	0%
Automatic [Step (6)]	26.1%	4.1%

pronunciations, and we investigate acoustic model estimation based on such improved transcriptions.

6. Improving the phonetic transcriptions used in acoustic model training

It is possible to gauge the quality of the phonetic transcription of Step (6) by comparing it with the manual phonetic transcription which is available for a portion of our Switchboard training corpus. The metric for this comparison is the string edit distance between the two phonetic transcriptions for each utterance. Time information about the phonetic segments is ignored in this alignment. The number of errors in the automatic transcriptions is the total number of insertions, deletions and substitutions. Table II presents this comparison for 1800 sentences (40 000 phones).

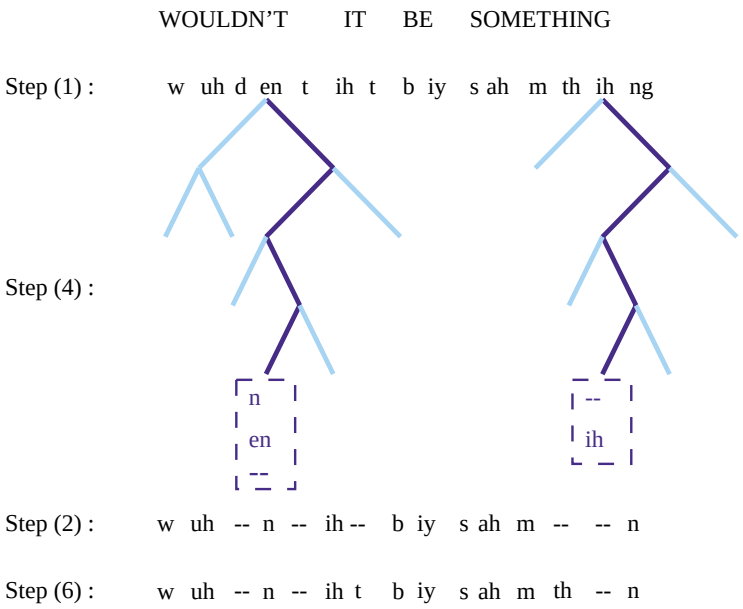

Figure 3. Phonemic [Step (1)], phonetic [Step (2)] and automatic [Step (6)] transcriptions and a decision-tree pronunciation model [Step (4)]. (The "- -" symbol corresponds to a deletion relative to the phonemic string.)

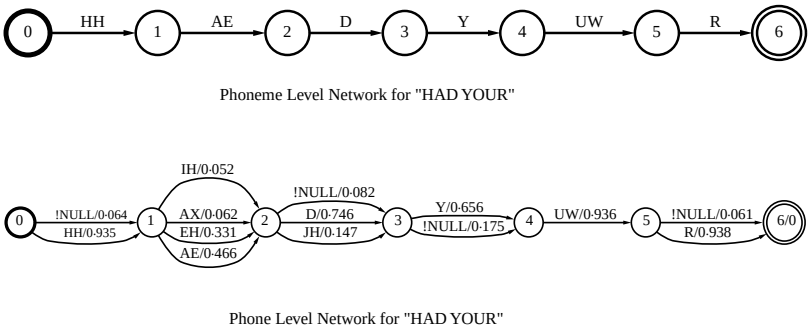

Figure 4. Phoneme- and phone-level networks for "HAD YOUR" (`!NULL` indicates a deletion). (Reproduced from the WS97 pronunciation modeling group final presentation by Michael Riley.)

Since the transcription resulting from Step (6) is more accurate, it may be expected to be better suited for acoustic model training than the canonical transcription. This leads to an additional step in the pronunciation modeling framework of Section 5.

(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

TABLE III. Failed attempts to further improve training transcriptions via adaptation

Adaptation method	Phone error rate	Deviation from baseforms
None	26.1%	4.1%
VTLN	26.0%	4.2%
MLLR	26.0%	4.0%

the *existing* acoustic models used earlier. The resulting phonetic transcription is then used in Steps (3)–(5) for pronunciation model estimation and recognition[2].

However, it was somewhat surprising to find that there was no improvement in recognition performance (38.9% WER) compared to the acoustic models trained on canonical baseforms!

While it is clear from Table II that the new models are reestimated on more accurate phonetic transcriptions than the canonical baseforms, it may be argued that a further improvement in phonetic accuracy of the training transcriptions is needed in order to improve recognition performance. We investigated four procedures to improve phonetic accuracy of the training transcriptions, as described below. The first three (adaptation, simpler models and cross-transcription) provide small improvements in training transcription accuracy and the fourth (bootstrapping) provides a significant improvement. We then reestimated the acoustic models from these improved transcriptions of the acoustic training corpus and used them for recognition in conjunction with a matched pronunciation model. We observed *no improvement in WER* from slight improvements in phonetic accuracy of the training transcriptions. Acoustic models trained on the fourth, significantly improved, training transcription, surprisingly, *degraded recognition WER*. These techniques are described in more detail below.

6.1. Speaker and channel adaptation

We first investigate whether standard speaker and channel adaptation techniques can be used to adjust the acoustic models used in Step (6) to obtain more accurate phonetic transcriptions. A variation of the Vocal Tract Length Normalization (VTLN) procedure proposed by Lee and Rose (1996) and Maximum Likelihood Linear Regression (MLLR) (Digalakis, Rtischev & Neumeyer, 1995; Leggetter & Woodland, 1995) are used to adjust the acoustic models before performing the retranscription in Step (6). The use of adaptation techniques leads to little change in transcription accuracy relative to the hand-labeled transcriptions (Table III). It also results in little change in transcription content as evidenced by the comparison of the three automatic transcription techniques in Table III. The new transcriptions remain fairly close to the original baseform transcriptions both before and after adaptation.

The results suggest the original hypothesis—that the phone transcription accuracy in Step (6) can be substantially improved within this framework—is incorrect; we conclude instead that the highly-parameterized acoustic models used here are well-tuned to match the acoustics to the PronLex baseforms on which they are trained so that only drastic mispronunciations can be discovered when using these models in the retranscription stage. Adaptation based on the training transcriptions simply reinforces the problem.

[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 subsequently in recognition in Step (5).

TABLE IV. Simpler acoustic models improve phonetic transcription

Acoustic model used in Step (6)	Phone error rate	Deviation from baseforms
12-Gaussian triphone models	26.1%	4.1%
8-Gaussian triphone models	25.7%	5.0%
1-Gaussian triphone models	25.5%	9.3%

TABLE V. Jack-knifing improves phonetic transcription (8-Gaussian triphone models)

Transcription technique	Phone error rate	Deviation from baseforms
Self-transcription	25.7%	5.0%
Cross-transcription	25.3%	8.1%

6.2. Simpler acoustic models

Since we suspect that the accuracy of phonetic transcriptions is being limited by the ability of the highly-parameterized acoustic models to match the realized acoustics to the canonical baseforms, we repeat the transcription of Step (6) using simpler acoustic models. In particular, we replace the 12-Gaussian mixtures in the HMM state output densities of the baseline system with 8-Gaussian mixtures or 1-Gaussian densities.

As seen in Table IV, the simpler acoustic models do slightly improve the phonetic accuracy of the training transcriptions.

6.3. Cross-transcription

The automatic transcription procedure of Step (6) may be hampered by the fact that the acoustic models used for transcription were trained on the *same* acoustics together with the canonical (baseform) transcription. A natural solution is to transcribe the training set using models trained on different data.

The 60-hour Switchboard training set is partitioned into two speaker disjoint gender-balanced 30 hour subsets and model sets trained on one half are used to phonetically transcribe the acoustics for the other 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 in that order to estimate and test a matched pronunciation model.

The phone recognition accuracy relative to the hand-labeled transcriptions improves slightly by the cross-transcription method as shown in Table V. 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 IV. Despite this, the "refined" transcriptions do not lead to any significant change in recognition performance.

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

6.4. 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 (ICSI portion of the corpus). We investigate this avenue as well.

Only a small portion (3.5 hours) of the acoustic training data has been transcribed at

TABLE VI. Using hand-labeled data to train acoustic models for improved phone transcription given the word transcription (451-utterance subset of the test set)

Transcription type	Models	Phone error rate	Deviation from baseforms
Dictionary baseforms	—	33.6%	0%
Automatic [Step (6)]	Standard	31.4%	3.9%
Automatic [Step (6)]	ICSI-models	26.6%	20.7%

TABLE VII. Comparison of word and phone error rates of 1-best recognition hypothesis under different acoustic and pronunciation models

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%

the phone level by human labelers. Due to this limitation, we estimate a set of context-independent phone models (henceforth called *ICSI-models*) using the hand-labeled portion of the training set. The limited amount of hand-labeled data has two unintended benefits. For one, most of the (60 hours of) speech to be transcribed is not used in model training, yielding some of the benefits of cross-transcription seen in Section 6.3. For another, the use of monophone models instead of triphones is another step in the direction of simpler acoustic models (for phonetic transcription) described in Section 6.2.

The automatic transcription of Step (6) is performed next, using the *ICSI-models* described above. This results in considerably more accurate phonetic training transcription (see Table VI). 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.

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 II–V 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 VI for the *ICSI-models* indicate that the transcriptions on which the *ICSI-bootstrap* models are trained are much more accurate than the baseforms or the transcriptions used in preceding sections.

Next we use the *ICSI-bootstrap* models for recognition. While the standard acoustic models (together with a pronunciation model) have a WER of 38.9%, the WER of the *ICSI-bootstrap* models turns out to be 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 analysed. In addition to the WER performance, the PER is measured against the hand transcriptions. It turns out (Table VII) 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%.

It is clear from these experiments that there is indeed considerable deviation from canon-

ical 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 believe that the inability to translate this (implicit) lower PER 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 more accurate phone recognition is difficult. Yet, the model for the acoustic realization of a phoneme must allow for the inherent variability. This very naturally leads to the quest in Section 7 for a modeling methodology in which *the HMM of a phoneme acquires the ability of the HMMs of alternate realizations to model the observed pronunciation variation, yet does not sacrifice its identity to the alternate realizations*.

7. Modeling pronunciation variability at the level of HMM states

We now present a new pronunciation model which 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).

To explain the mechanism of state-level pronunciation modeling we will use `typewriter` fonts to denote pronunciations of words, both canonical and alternate. For example, the word HAD has a canonical pronunciation `hh ae d` but it may sometimes be realized as `hh eh d`, which happens to be the canonical pronunciation of the word HEAD. The sketch at the top of Figure 5 illustrates how this alternative will be represented at the phone level, and the sketch in the middle shows how a context-independent HMM system will permit this alternative in a recognition network.

The SLPM deviates from these methods as illustrated by the sketch at the bottom of Figure 5. Rather than letting the phoneme `ae` be realized as an alternate phone `eh`, the HMM states of the acoustic model of the phoneme `ae` are instead allowed to mix-in the output densities of the HMM states of the acoustic model of the alternate realization `eh`. Thus the acoustic model of a phoneme `ae` has the canonical and alternate realizations (`ae` and `eh`) represented by different sets of mixture components in one set of HMM states.

In a system that uses context-dependent phone models, a pronunciation change (`ae`→`eh`) also affects the HMMs selected for the neighboring phones, as illustrated by the sketch in the middle of Figure 6. The mechanism used by the SLPM to accommodate pronunciation changes in a phoneme's context is to allow the states of the context-dependent HMM of the phoneme `d`, namely `ae-d+sil`, to mix-in the output densities of the states of `eh-d+sil`, the context-dependent HMM for the alternative pronunciation. This is illustrated by the sketch at the bottom of Figure 6.

In our system, we use three-state left-to-right HMMs to model triphones. To reduce model complexity, all triphone states of a phone are clustered into a manageable number of states using a top-down decision-tree procedure (Odell, 1995; Young *et al.*, 1995). A separate clustering tree is grown for each of the three states in a HMM. The detailed effect of the SLPM on such a system is illustrated in Figure 7. Each HMM state of the triphone `hh-ae+d` shares output densities with the corresponding HMM state of the triphone `hh-eh+d` to accommodate a pronunciation change `ae`→`eh` in the phoneme, as illustrated by the sketch on the left in Figure 7. Similarly, each HMM state of the triphone `ae-d+sil` shares output densities with the corresponding HMM state of the triphone `eh-d+sil` to accommodate a pronunciation change `ae`→`eh` in the left context. This is illustrated by the sketch on the right in Figure 7.

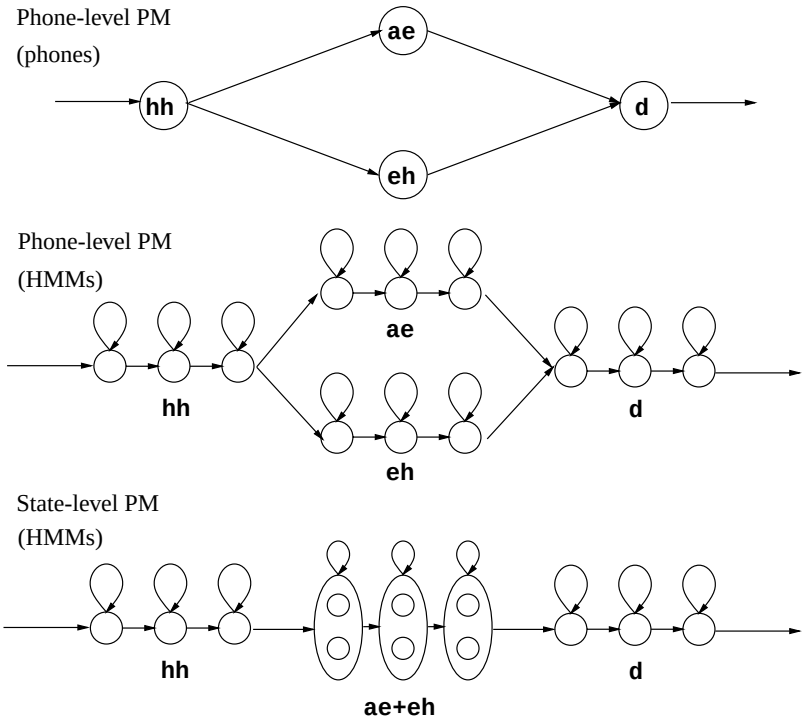

Figure 5. The effect of allowing phoneme `ae` to be realized as phone `eh`: context-independent models.

Note that each HMM state of each triphone, say of phoneme x, shares output densities with two *kinds* of states: (i) corresponding states of a phoneme y, caused by a pronunciation change $x \rightarrow y$, and (ii) corresponding states of other triphones of the phoneme x, caused by pronunciation changes in the triphone context $C(x)$ of x. This overall effect of the SLPM on a HMM state is illustrated in Figure 8. The trees in Figures 7 and 8 are state-clustering trees, *not* the pronunciation prediction trees illustrated in Figure 3.

7.1. A state-level pronunciation model

Having introduced the philosophy of state-level pronunciation modeling, we now present a detailed recipe for deriving such a model. We describe a procedure for identifying, for each HMM state in our system, the set of other HMM states with which it shares output densities. We then formulate the implementation of this sharing when each individual output density is a mixture of Gaussians.

alignment. Starting with a canonical phonemic transcription of the acoustic training corpus [Section 5, Step (1)] and its surface-form phonetic transcription [Section 5, Step (6)], we obtain an alignment between the two corresponding sequences of (context-dependent) HMM states. We have examined two ways of obtaining a state-to-state alignment: (i) we use the phoneme-to-phone alignment of Section 5, Step (3), to infer a state-to-state alignment; and (ii) we first obtain a Viterbi alignment of each HMM state sequence with the (common) acoustic signal and use the two resulting sets

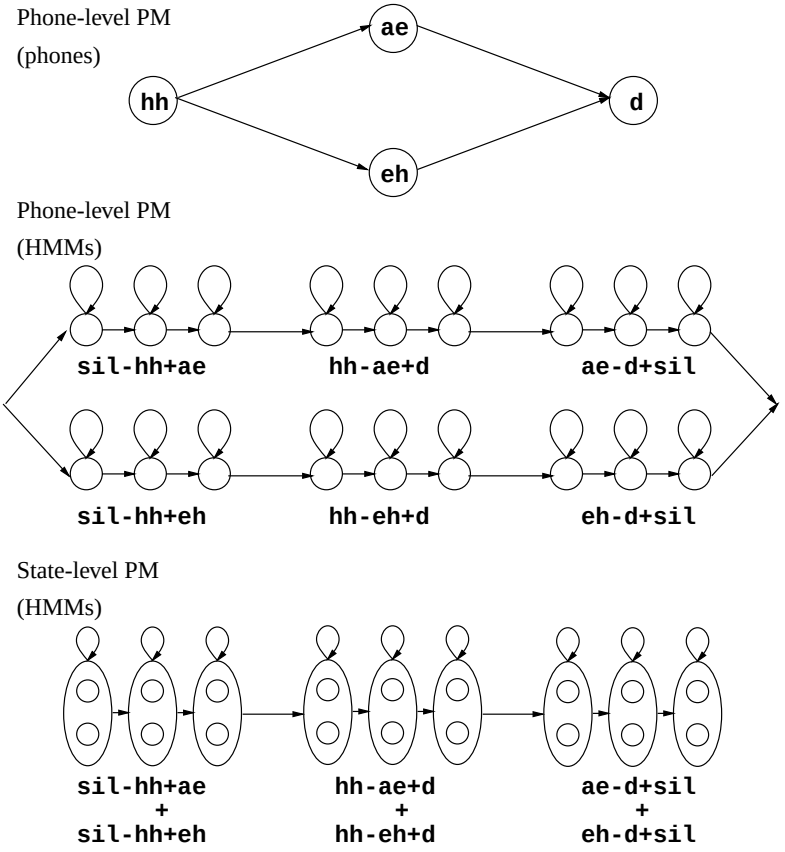

Figure 6. The effect of allowing a phoneme `ae` to be realized as a phone `eh`: context-dependent models.

of frame-to-state alignments to infer a state-to-state alignment. While these methods result in slightly different alignments, the eventual SLPMs obtained by the two methods were observed to have identical recognition WER. This suggests that our recipe is insensitive to the details of the alignment procedure. Therefore, we will present experimental results only for the first method of obtaining state-to-state alignment.

counting. Using the state-to-state alignment obtained above, compute the relative frequency with which a state b in the sequence of states for the baseform transcription is aligned to a state s in the sequence of states for the surface form transcription.

$$\text{Freq}(s|b) = \frac{\text{Count}(s, b)}{\text{Count}(b)}.$$

estimation. Estimate the probability, $P(s|b)$, of a state b being aligned to a state s by using the relative frequency and filtering out infrequent or unlikely alternate realizations. In particular, we discard all (b, s) pairs with counts less than 10, and ignore all realizations s of b with frequencies less than 5%.

$$\begin{array}{ll} P(s|b) = 0 & \text{if Count}(s, b) < \text{threshold}_1 (= 10) \\ P(s|b) = 0 & \text{if Freq}(s|b) < \text{threshold}_2 (= 0.05) \\ P(s|b) \propto \text{Freq}(s|b) & \text{otherwise.} \end{array}$$

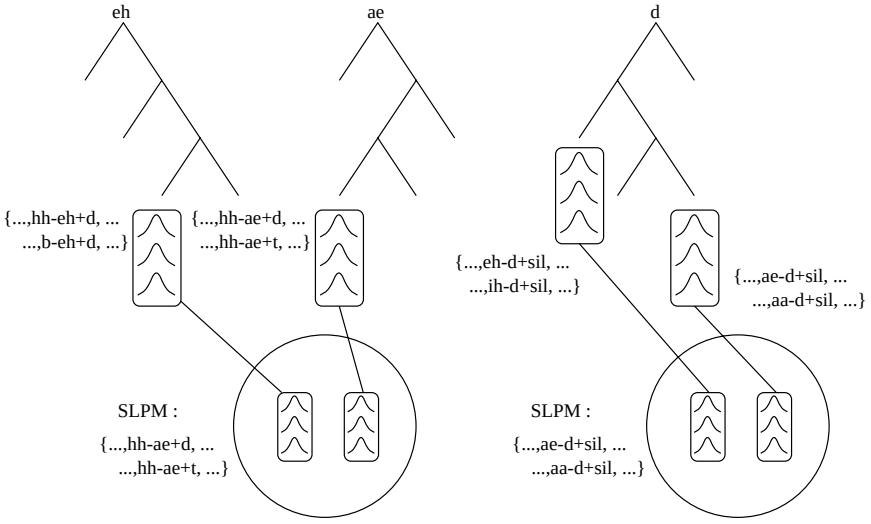

Figure 7. Sharing Gaussian mixtures among different HMM states to accommodate the pronunciation change `ae`→`eh`.

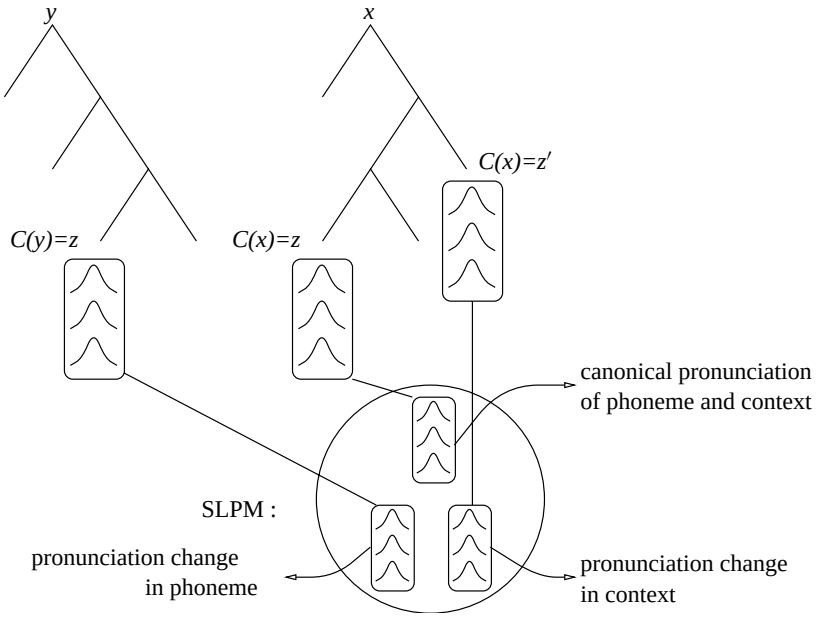

Figure 8. Overall effect of sharing Gaussian mixtures among different tree-clustered HMM states to accommodate pronunciation changes $x \rightarrow y$ and $z \rightarrow z'$ where z and z' are contexts of x.

modification. Replace the output distribution of each state b with a mixture of the output distributions of its alternate realizations s. Use $P(s|b)$ as the mixture weights. Note that the output distribution of a state s in the original system may be a mixture component in the output distribution of more than one state b in the new system. However, unlike a traditional *tied-mixture* (Bellegarda & Nahamoo, 1990) or *semi-continuous* (Huang & Jack, 1989) system, the tying of mixture components is governed by a pronunciation model. The new output distribution of state b is given by

$$P'(o|b) = \sum_{s:P(s|b)>0} P(s|b)P(o|s)$$

where $P(o|s)$ is the output distribution of state s in the original system.
The state output densities in our system are themselves mixtures of Gaussians

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

where i indexes *all* the Gaussian densities used in the system, $G(s)$ is the subset of densities which are used by the state s and $w_{i,s}$ is the mixture weight of the density i as used in state s. For such a system, this step yields

$$\begin{aligned} P'(o|b) &= \sum_{s:P(s|b)>0} P(s|b) \sum_{i \in G(s)} w_{i,s}\mathcal{N}(o; \mu_i, \Sigma_i) \\ &= \sum_{i \in G'(b)} w'_{i,b}\mathcal{N}(o; \mu_i, \Sigma_i) \end{aligned} \tag{1}$$

where

$$w'_{i,b} = \sum_{s:P(s|b)>0} w_{i,s}P(s|b),$$

$$G'(b) = \bigcup_{s:P(s|b)>0} G(s).$$

A further simplification is possible if in the original system each Gaussian is associated with only one state, i.e. $w_{i,s} > 0$ for only one $s = s(i)$. This is true for our system, which allows for expressing the new mixture weights as

$$w'_{i,b} = w_{i,s(i)}P(s(i)|b). \tag{2}$$

reestimation. Further train the resulting "tied-mixture"-like acoustic models. Note that it is easy to separate the factors $P(s|b)$ and $w_{i,s}$ in $w'_{i,b}$ of Equation (2). However, treating $w'_{i,b}$ as a single parameter allows for further training of the models with the Baum–Welch algorithm without any need for a modification of the training software. We therefore initialize $w'_{i,b}$ as described in Equation (2) but reestimate it as a free variable. Retaining the factorization is possible and requires a slightly more involved training algorithm described by Luo (1999) and used recently by Hain and Woodland (1999*b*).

The SLPM developed above is used in a recognition experiment on the Switchboard corpus. Table VIII shows that, just as for the decision-tree pronunciation model, the SLPM results in a small but significant reduction in WER over the baseline system.

It may be noted from Equation (1) that the Gaussian densities indexed by i are not duplicated before reestimation, but are shared among states. The only additional parameters introduced are the mixture weights, $w'_{i,b}$. This increases the number of free parameters in the

TABLE VIII. Recognition performance of the SLPM

Pronunciation model	WER
None (PronLex dictionary)	39.4%
Decision-tree pronunciation model	38.9%
State-level pronunciation model	38.8%

acoustic models by less than 0.5%. The amount of data required to reestimate the system is therefore comparable to that for the baseline system.

The mechanism of sharing output densities among HMM states is inspired by the doctoral dissertation of Luo (1999). Motivated by the desire to obtain more reliable estimates of the output densities of a collection of HMM states, some of which see plenty of training data while others see very little, Luo replaces the output density of each state with a mixture of its original output density and the output densities of a fixed number of "similar" states. The choice of these sibling states is limited to states within the same triphone-clustering tree and similarity of states is determined by a statistical distance between the output densities. Luo calls this procedure *soft clustering*.

The motivation for the work here is clearly different, namely accommodating pronunciation variability. As a consequence, we select a state s in the original system for contributing densities to a state b in the new system if our pronunciation model suggests that $P(s|b) > 0$. This results in HMM states of triphones of distinct phones sharing output densities. The number of states s which contribute to the output density of a state b is also not fixed, it varies with the pronunciation variability of the (tri)phone of which b is a state.

The SLPM has the following advantages over the decision-tree pronunciation model of Riley *et al.* (1999):

- Canonical (phonemic) transcriptions [Section 5, Step (1)] 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, instances in which the acoustic realization matches an alternate phone [Section 5, Step (6)] better than the canonical one will be used by the Baum–Welch procedure to update the densities of the alternate phone instead of the canonical one.
- The dictionary need not be expanded to include alternate pronunciations, an important consideration for recognition speed. The number of Gaussian computations, however, increases due to the larger number of Gaussian components associated with each state. The baseline system has 12 Gaussians per state. The number of Gaussians per state for the SLPM varies and, on average, there are 14.5 Gaussians per state.

We argued in Section 6.4 that the phone-level decision-tree pronunciation model may end up introducing excessive lexical confusion by, for instance, permitting HAD to have a pronunciation `hh eh d`, which is the canonical pronunciation of HEAD. The SLPM contains only the canonical pronunciation `hh ae d` of HAD, and HMM states of the triphones of this pronunciation share output densities with the states in the alternate pronunciation, e.g. `hh-ae+d` with `hh-eh+d`. This leads to several differences in the two models. To see the main difference, assume that HEAD is always pronounced as `hh eh d` while HAD has two alternate pronunciations, `hh ae d` and `hh eh d`. To simplify the illustration, let HAD and HEAD have identical language model probabilities in some context. In this context, consider an acoustic realization which has a few acoustic frames with high likelihood under the out-

put densities of `hh-ae+d` and many frames with high likelihood under the output densities of `hh-eh+d`. The two pronunciation models handle this realization differently. Under the phone-level decision-tree pronunciation model, the phone sequence `hh ae d` obtains a very small acoustic score relative to the phone sequence `hh eh d`. Since both HAD and HEAD have a pronunciation `hh eh d`, the lexical choice is determined by the pronunciation model to be HEAD. Under the SLPM, such an acoustic realization may obtains a higher score under the phone sequence `hh ae d` than `hh eh d`, due to the fact that the *modified* HMM for `hh-ae+d` utilizes the output densities of `hh-eh+d` but not vice versa. Since HAD only admits the pronunciation `hh ae d`, and HEAD `hh eh d`, there is no confusion at the lexical level, and the word HAD is the winner. This is not to say that HAD is the better of the two choices but that the two pronunciation models make different choices.

7.2. Introducing more accurate densities to model the surface-form realizations

The essential idea of the SLPM described above is to augment the output density of an HMM state of a phoneme so as to model alternate surface-form realizations. However, all output densities of the system described above were first estimated from the canonical transcriptions of the acoustic training set. For example, in Figure 8, when an HMM state b needs to account for a realization s, an output density of s *estimated from canonical transcriptions* is used. An alternative is to augment the output density of b with the output density of the state corresponding to s in a HMM system *trained on the more accurate surface-form transcriptions* such as the one described in Section 6.4. Recall that the *ICSI-bootstrap* models of Section 6.4 improve phone accuracy on the test set.

Recall further that phonetic transcriptions which are considerably more accurate than the canonical baseforms (see Table VI) were used for estimating the ICSI-bootstrap models. Now, there is a sequence of triphone states, $\{s'\}$, from the ICSI-bootstrap models corresponding to this (more accurate) phonetic transcription. Analogous to the **alignment** step in the SLPM recipe described in Section 7.1, we obtain an alignment of this state sequence with the state sequence corresponding to the canonical transcription. Analogous to the **counting** and **estimation** steps, we then estimate the probability $Q(s'|b)$ that a state s' in the ICSI-bootstrap models aligns with a state b in the canonical transcription.

The main modification to the SLPM recipe of Section 7.1 is in the **modification** step.

alternate modification. Replace the output distribution of each state b with a mixture of the output distributions of its alternate realizations s, with $P(s|b) > 0$, in the original acoustic models *and* the output distribution of alternate realizations s', with $Q(s'|b) > 0$, in the ICSI-bootstrap acoustic models[3]. This is illustrated in Figure 9.

Further training of the models (cf. **reestimation** step in Section 7.1) is done by first training the mixture weights and transition probabilities followed by training the whole model.

The results in Table IX 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; its WER on the test set is 39.7%, worse than the 12 Gaussians-per-state baseline.

In another effort to make a fair comparison by keeping the number of parameters in the

[3]Note that this does substantially increase the number of acoustic model parameters in the system.

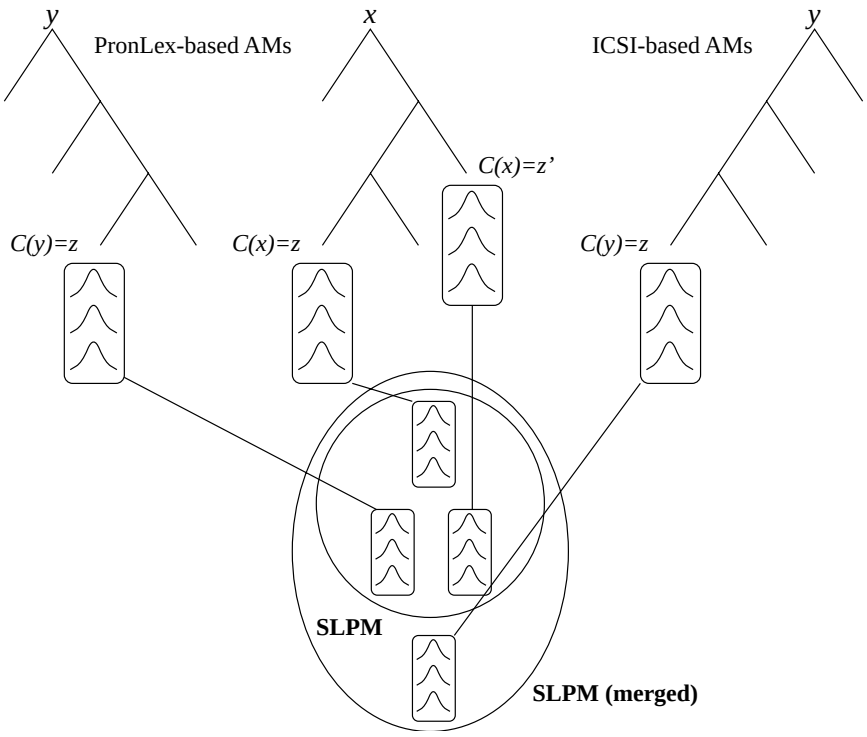

Figure 9. Sharing Gaussian mixtures in the "merged" SLPM, to be contrasted with the SLPM in Figure 8.

TABLE IX. Recognition performance of the "merged" SLPM

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%

SLPM comparable to the baseline models, two sets of acoustic models one corresponding to the baseline system and another corresponding to the ICSI-bootstrap models, each with a smaller number of mixture components (six per state), are merged. This results in a merged system with roughly the same number of parameters as our baseline. This system has a WER of 38.3%, which is still substantially better than the decision-tree pronunciation model.

As a final check, we have validated our results on a different Switchboard test set (WS97 eval set), consisting of 882 utterances. This test set was not used at all until all other experiments described above were completed. The results are given in Table X.

Table X also provides a highly abbreviated summary of the work presented in this paper. Compared to a system that uses a dictionary with canonical pronunciations, traditional (decision tree) pronunciation models provide modest reductions in WER. The improvement in performance due to the SLPM described in Section 7.1 is comparable, if not better. By incor-

TABLE X. Recognition performance on a blind-test (WS97 eval set)

Pronunciation model	Acoustic model	WER
PronLex (baseline)	PronLex based	36.7%
Phone-level PM	PronLex based	36.5%
State-level PM	PronLex based	36.2%
State-level PM	Merged, no training	35.8%
State-level PM	Merged, further training	35.1%

porating output densities from a system trained on surface-forms, the SLPM of Section 7.2 is able to significantly reduce the WER of the system.

8. Conclusion

It has been shown that conversational speech exhibits a high degree of pronunciation variation. Using models that enhance the pronunciation dictionary to accommodate this variation gives a moderate improvement in system performance. Although the idea of using a pronunciation model to obtain better phonetic transcriptions for acoustic model training is appealing, attempts in doing so have not yielded any improvement in recognition performance in the past.

We have shown here that starting with acoustic models trained directly on a phonetically labeled corpus together with a very rich pronouncing dictionary is a good way to obtain fairly accurate phonetic transcriptions of a larger corpus for acoustic HMM training.

We have demonstrated that acoustic models trained on these improved phonetic transcriptions are indeed more accurate phonetic models. However, using such a rich dictionary for recognition degrades accuracy, possibly due to an undesirable degree of lexical confusion.

It is more fruitful to use the improved training transcriptions and the HMMs estimated from them in a new method of modeling pronunciation variation implicitly in HMMs for the phones instead of explicit modeling in the dictionary. This new implicit method (SLPM) achieves a significant reduction in WER on the Switchboard corpus as seen on two independent test sets.

The authors thank Entropic Cambridge Research Laboratory, U.K., for providing the lattice generation tools used in these experiments, Michael Riley of AT&T Laboratories for providing FSM tools required to construct decision-tree pronunciation models and manipulate the lattices for rescoring, and William Byrne of Johns Hopkins University for suggesting the cross-transcription experiment of Section 6.3. This research was partially supported by the U.S. National Science Foundation (Grant No. IRI9714196) and the U.S. Department of Defence (Contract No. MDA90499C3525).

References

Bellegarda, J. R. & Nahamoo, D. (1990). Tied mixture continuous parameter modeling for speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **38**, 2033–2045.

Bernstein, J., Baldwin, G., Cohen, M., Murveit, H. & Weintraub, M. (1986). Phonological studies for speech recognition. In *DARPA Speech Recognition Workshop*, pp. 41–48.

Byrne, W., Finke, M., Khudanpur, S., McDonough, J., Nock, H., Riley, M., Saraclar, M., Wooters, C. & Zavaliagkos, G. (1997). Pronunciation modelling for conversational speech recognition: A status report from WS97. In *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings (ASRU)*, Santa Barbara, CA, USA, pp. 26–33.

Byrne, W., Finke, M., Khudanpur, S., McDonough, J., Nock, H., Riley, M., Saraclar, M., Wooters, C. & Zavaliagkos, G. (1998). Pronunciation modelling using a hand-labelled corpus for conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, USA, pp. 313–316.

Chen, F. (1990). Identification of contextual factors for pronunciation networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 753–756.

Cohen, M. (1989). *Phonological structures for speech recognition.* Ph.D. Thesis. University of California, Berkeley.

Digalakis, V. V., Rtischev, D. & Neumeyer, L. G. (1995). Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, **3**, 357–366.

Eide, E. (1999). Automatic modeling of pronunciation variations. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary, pp. 451–454.

Eide, E., Gish, H., Jeanrenaud, P. & Mielke, A. (1995). Understanding and improving speech recognition performance through the use of diagnostic tools. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, MI, pp. 221–224.

Finke, M., Fritsch, J., Koll, D. & Waibel, A. (1999). Modeling and efficient decoding of large vocabulary conversational speech. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary, pp. 467–470.

Finke, M. & Waibel, A. (1997). Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pp. 2379–2382.

Fosler, E., Weintraub, M., Wegmann, S., Kao, Y.-H., Khudanpur, S., Galles, C. & Saraclar, M. (1996). Automatic learning of word pronunciation from data. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. S28–S29 (addendum).

Fosler-Lussier, E. (1999). Multi-level decision trees for static and dynamic pronunciation models. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary, pp. 463–466.

Giachin, E., Rosenberg, A. & Lee, C. (1990). Word juncture modeling using phonological rules for HMM-based continuous speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 737–740.

Godfrey, J., Holliman, E. & McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 517–520, Available at `http://www.ldc.upenn.edu/`.

Greenberg, S. The switchboard transcription project. Technical Report, 1996 LVCSR Summer Workshop. `http://www.icsi.berkeley.edu/real/stp/`.

Hain, T. & Woodland, P. (1999*a*). Recent experiments with the CU-HTK Hub5 system. *10th Hub-5 Conversational Speech Understanding Workshop*.

Hain, T. & Woodland, P. C. (1999*b*). Dynamic HMM selection for continuous speech recognition. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Budapest, Hungary, pp. 1327–1330.

Huang, X. D. & Jack, M. A. (1989). Semicontinuous hidden Markov models for speech signals. *Computer Speech and Language*, **3**, 239–251.

Lee, L. & Rose, R. (1996). Speaker normalization using efficient frequency warping procedures. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, GA, pp. 353–356.

Leggetter, C. J. & Woodland, P. C. (1995). Speaker adaptation of continuous density HMMs using multivariate linear regression. *Computer Speech and Language*, **9**, 171–185.

Lucassen, J. M. & Mercer, R. L. (1984). An information theoretic approach to the automatic determination of phonemic baseforms. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Diego, CA, pp. 42.5.1–42.5.4.

Luo, X. (1999). *Balancing Model Resolution and Generalizability in Large Vocabulary Continuous Speech Recognition.* PhD Thesis. The Johns Hopkins University, Baltimore, MD.

McAllaster, D., Gillick, L., Scattone, F. & Newman, M. (1998). Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, pp. 1847–1850.

Mohri, M., Pereira, F. C. N. & Riley, M. (2000). The design principles of a weighted finite state transducer library. *Theoretical Computer Science*, **231**, 17–32, Available from `http://www.research.att.com/sw/tools/fsm/`.

Odell, J. (1995). *The Use of Context in Large Vocabulary Speech Recognition.* PhD Thesis. Cambridge University Engineering Department.

Peskin, B., Newman, M., McAllaster, D., Nagesha, V., Richards, H., Wegmann, S., Hunt, M. & Gillick, L. (1999). Improvements in recognition of conversational telephone speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 53–56.

PronLex, COMLEX English Pronunciation, Available from `http://www.ldc.upenn.edu/`.

Riley, M. (1991). A statistical model for generating pronunciation networks. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 737–740.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C. & Zavaliagkos, G. (1999). Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, **29**, 209–224.

Riley, M. & Ljolje, A. (1995). Automatic generation of detailed pronunciation lexicons. *Automatic Speech and Speaker Recognition : Advanced Topics*, chapter 12, pp. 285–302. Kluwer Academic Press.

Sloboda, T. & Waibel, A. (1996). Dictionary learning for spontaneous speech recognition. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, pp. 2328–2331.

Strik, H. & Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication*, **29**, 225–246.

Tajchman, G., Fosler, E. & Jurafsky, D. (1995). Building multiple pronunciation models for novel words using exploratory computational phonology. *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Madrid, Spain, pp. 2247–2250.

Wakita, Y., Singer, H. & Sagisaka, Y. (1999). Multiple pronunciation dictionary using HMM-state confusion characteristics. *Computer Speech and Language*, **13**, 143–153.

Weintraub, M., Taussig, K., Hunicke-Smith, K. & Snodgrass, A. (1996). Effect of speaking style on LVCSR performance. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. S16–S19 (addendum).

Wooters, C. & Stolcke, A. (1994). Multiple pronunciation lexical modeling in a speaker independent speech understanding system. *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 1363–1366.

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

(Received 8 September 1999 and accepted for publication 11 February 2000)