Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Computer Speech and Language xxx (2003) xxx–xxx

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Pronunciation change in conversational speech and its implications for automatic speech recognition ☆

Murat Saraçlar [a,*], Sanjeev Khudanpur [b]

[a] *AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932, USA*
[b] *Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA*

Received 24 March 2003; received in revised form 16 September 2003; accepted 19 September 2003

Abstract

Pronunciations in spontaneous speech differ significantly from citation form and pronunciation modeling for automatic speech recognition has received considerable attention in the last few years. Most methods describe alternate pronunciations of a word using multiple entries in a dictionary or using a network of phones, assuming implicitly that a deviation from the canonical pronunciation results in a "complete" change as described by the alternate pronunciation. We investigate this implicit assumption about pronunciation change in conversational speech and demonstrate here that in most cases, the change is only partial; a phone is not completely deleted or substituted by another phone but is modified only partially. Evidence supporting this conclusion comes from the three-way analysis of features extracted from the acoustic signal for use in a speech recognition system, canonical pronunciations from a dictionary, and careful phonetic transcriptions produced by human labelers. Most often, when a deviation from the canonical pronunciation is marked, neither the canonical nor the manually labeled phones represent the actual acoustics adequately. Further analysis of the manual phonetic transcription reveals a significant number (>20%) of instances where even human labelers disagree on the identity of the surface-form. In light of this evidence, two methods are suggested for accommodating such partial pronunciation change in the automatic recognition of spontaneous speech and experimental results are presented for each method.

☆ This work was supported by the US Department of Defense via Contract No. MDA90499C3525.
* Corresponding author.
E-mail addresses: murat@research.att.com (M. Saraçlar), khudanpur@jhu.edu (S. Khudanpur).

0885-2308/$ - see front matter
doi:10.1016/j.csl.2003.09.005

1. Introduction

Most state of the art systems for automatic speech recognition (ASR), especially those for large vocabulary spontaneous speech, use acoustic models of phone-sized segments of speech as their building blocks. The phonetic representation of a word is usually obtained by looking it up in a hand-crafted pronunciation dictionary or deriving it from some set of rules, e.g., those written for a text to speech system. Having obtained such a canonical pronunciation, the acoustic model for the word is then composed by "stringing together" the (usually context-dependent) acoustic models of the corresponding phones. It therefore should not come as a surprise that such systems encounter considerable difficulties in recognizing spontaneous speech, where pronunciations of words often do not conform to citation form.

Recent speech recognition literature has seen numerous efforts aimed at addressing this problem under the banner of *pronunciation modeling*. Most of the proposed solutions aim one way or another to hypothesize plausible alternate pronunciations of a word, i.e., pronunciations not encountered in a dictionary, which are then used for (compositionally) creating alternate acoustic models for the word. These acoustic models therefore implicitly rely on the existence of accurate phonemic representations of the alternate pronunciations of words. Examples of such efforts include Finke and Waibel (1997), Ma et al. (1998), Fosler-Lussier (1999), Stolcke et al. (2000), Jurafsky et al. (2001), etc.

We show in this paper that for spontaneous speech in particular, and perhaps more generally, the degree of deviation from the canonical pronunciation varies on a continuum. Most of the time the deviation is not large enough to be clearly identifiable at the symbolic (phonemic) level. In other words, a phone is not completely deleted or substituted by another phone, but it is modified only partially and the effects of this modification can be found in its acoustic environment. Evidence supporting this conclusion is presented here via a three-way analysis of

1. acoustic features extracted from the speech signal for use in an ASR system,
2. the canonical pronunciations derived from a dictionary, and
3. an actual phonetic transcription produced by human labelers.

It is shown that most often, when a deviation from the canonical pronunciation is marked by the labelers, neither the canonical nor the manually labeled phones represent the acoustics well.

Corroborating evidence for these partial deviations is also found in the manual phonetic transcriptions alone. An analysis of a portion of the Switchboard corpus labeled independently by two transcribers at the phonetic level reveals that the disagreement between human labelers is quite high. We suggest that when the deviation from the canonical pronunciation is neither small nor large, but at an intermediate level, the transcribers, who are forced to use a categorical label from the limited phonetic inventory at their disposal, may end up choosing different labels for a surface-form. These points of inter-labeler disagreement, which correlate very well with the difficulty of *automatic* phonetic transcription, are consistent with partial pronunciation change.

The degree of deviation from the canonical pronunciation, of course, has implications for the strategy to be used to model pronunciation variation in ASR. If the deviations are small enough, then they can be absorbed within the acoustic model. If the deviations are large enough that they can be identified at the symbolic level as the standard realization of a different phoneme, then pronunciation modeling at the phone level, as traditionally done, may be adequate. If, on the

other hand, a majority of the deviations are at an intermediate level, then alternate techniques need to be developed. This paper discusses two such techniques.

The rest of this paper is organized as follows. In Section 2, we present acoustic evidence for partial pronunciation change in conversational speech and further characterize the resulting acoustic realization as spectrally (Section 2.1) or temporally (Section 2.3) intermediate between the canonical and an "alternate" phonetic realization. The impact or consequence of this partial change in the task of manual and automatic phonetic transcription is presented in Section 3, where we demonstrate an automatic phonetic transcription procedure whose accuracy rivals that of human labelers. Pronunciation modeling techniques that take cognizance of this phenomenon of partial pronunciation change are presented in Section 4, along with experimental ASR results. We conclude with a summary of the main points in Section 5.

1.1. The Switchboard corpus

We use the Switchboard corpus (cf Godfrey et al., 1992), a collection of spontaneous telephone conversations between American English speakers, to study the nature of pronunciation change. Switchboard contains a large amount of orthographically transcribed speech, and is a well-studied corpus for automatic speech recognition, making it possible to assess the impact of pronunciation modeling ideas in a large vocabulary speech recognition system. We use a subset of approximately 60 h of speech from this corpus to train acoustic models for the ASR system, and a disjoint subset of approximately 2 h to measure the accuracy of the ASR system. There are no common speakers between the two parts.

A portion of Switchboard (approximately 4 h, comprising $\sim$100K phones) has been phonetically labeled with great care by Greenberg (1998), and this is what makes our investigation possible. Approximately 30 min of this phonetically transcribed material falls within our test partition of Switchboard, making it possible to also measure the phonetic recognition accuracy of various acoustic models, and provides an important diagnostic tool.

Unless mentioned otherwise, all word error rate (WER) results in this paper are on the 2-h test set mentioned above, and all phone error rates (PERs) are on the phonetically transcribed 30-min subset of the 2-h test set. An important exception is a 3-min portion of the phonetically labeled corpus ($\sim$2000 phones) that has been phonetically labeled independently by two transcribers. This doubly labeled portion makes it possible, as is done in Section 3, to measure interlabeler agreement on this task, and to further characterize the phonetic accuracy of the acoustic models used by the ASR system.

2. Acoustic analysis of pronunciation variation

We begin by presenting evidence that pronunciation variation leads to surface-form realizations of a phoneme which span a continuum. One end of this continuum is the canonical realization of the phoneme while the other end is what may uncontroversially be called a *complete* pronunciation change – to a realization which is reliably identified by a human labeler. We demonstrate that there are several realizations which lie between these extremes and the categorical label the transcriber ought to ascribe to them is unclear. In order to understand the

nature of this kind of variation, we need to investigate the three-way relationship between the acoustics, the canonical pronunciation or *base-form*, and the surface-form representation whenever a pronunciation change occurs.

Consider an occurrence of the word HAD which has the base-form /hh ae dd/ and is manually labeled as having been realized as the surface-form [hh eh dd]. In this example, /ae/ is realized as [eh] forming the base-form/surface-form pair (ae: eh). We use the notation /b/ to denote base-form phonemes, [s] to denote surface-form phones, and (b: s) to denote base-form/surface-form pairs.

What do the acoustics of this pair look like?

If the acoustics are sufficiently similar to those of an /eh/ realized elsewhere as an [eh] (i.e., to those of (eh: eh), according to our notation), this can be considered an instance of *complete pronunciation change*, otherwise this is a case of *partial pronunciation change*.

How should this pair be modeled in an ASR system?

Complete pronunciation change may perhaps be dealt with by traditional models of pronunciation variation, e.g., by adding /hh eh dd/ as a second dictionary entry [1] for HAD, whereas partial pronunciation change requires other solutions.

In order to answer these questions we treat the base-form/surface-form pair, e.g., (ae: eh), as a unit and analyze the acoustics of such units. We begin our analysis by studying the phonetically labeled portion of our acoustic training set.

- The speech signal is processed using a standard acoustic "front-end" used by our ASR system (cf., e.g., Young et al., 1995, for details). This results in the extraction of 13 MF-PLP coefficients every 10 ms, each covering a 25 ms analysis window. First and second order differences (Δ and $\Delta\Delta$) are appended, yielding a 39-dimensional "observation." The mean-value of the observations over an entire conversation-side is subtracted from each observation, a procedure sometimes referred to as *cepstral mean removal*. The resulting sequence of 39-dimensional acoustic observations is denoted by $\mathbf{A}$.
- The base-form transcriptions $\hat{\mathbf{B}}$, obtained by dictionary lookup, and the surface-form transcriptions $\hat{\mathbf{S}}$, i.e., the manual phonetic labels, are aligned to obtain "pair transcriptions" $\widehat{\mathbf{BS}}$ of the acoustics $\mathbf{A}$.
- Three sets of context independent acoustic phonetic models, as listed below, are then estimated from these sets of transcriptions via standard techniques.
  1. $P_{\mathbf{B}}(\mathbf{A}|\hat{\mathbf{B}})$: estimated from the base-form transcriptions $\hat{\mathbf{B}}$.
  2. $P_{\mathbf{S}}(\mathbf{A}|\hat{\mathbf{S}})$: estimated from the surface-form transcriptions $\hat{\mathbf{S}}$.
  3. $P_{\mathbf{B:S}}(\mathbf{A}|\widehat{\mathbf{BS}})$: estimated from the pair transcriptions $\widehat{\mathbf{BS}}$.

Note that $P_{\mathbf{B}}(\cdot|\cdot)$ and $P_{\mathbf{S}}(\cdot|\cdot)$ have about 50 hidden Markov models (HMMs), one for each phoneme or phone, whereas $P_{\mathbf{B:S}}(\cdot|\cdot)$ has roughly 500 HMMs, out of the 2500 possible phoneme–phone pairs, since not all possible pairs are seen in our phonetically labeled corpus.

In each of the three acoustic models, a phoneme, phone or phone-pair is represented by a three-state left-to-right HMM. The probability density function (pdf) of the acoustics "emitted" by each HMM state is assumed to be a *single* 39-variate Gaussian with a diagonal covariance matrix. The mean of the Gaussian pdf corresponding to each state of a particular phone is therefore repre-

[1] Note that this adds confusion at the lexical level: the word HEAD also has the pronunciation /hh eh dd/.

sentative of the acoustics of that segment of that phone. These are the models used throughout the investigation of the nature of pronunciation variation reported in Sections 2.1–2.3.

2.1. Acoustics of the alternate realizations

Our analysis begins by visualizing how the average acoustic features corresponding to an instance of a base-form phoneme /b/ that is realized as a surface-form phone [s] compare to the average of

(i) all realizations of the base-form phoneme /b/, no matter what their realized surface-forms are, and
(ii) all base-form phonemes that are realized as the surface-form phone [s], no matter what the base-form phoneme is.

Note that since we estimate single Gaussian output pdfs for our acoustic models $P_{\mathbf{B:S}}(\mathbf{A}|(\mathtt{b}:\mathtt{s}))$, the model-mean $\mu_{\mathrm{B:S}}((\mathtt{b}:\mathtt{s}))$ may also be interpreted as a "typical" acoustic frame when a phoneme /b/ is realized as a phone [s]. Similarly, $\mu_{\mathrm{B}}(/\mathtt{b}/)$ may be considered the average of item (i) above, and $\mu_{\mathrm{S}}([\mathtt{s}])$ that of item (ii). We therefore focus attention on the relative location of $\mu_{\mathrm{B:S}}((\mathtt{b}:\mathtt{s}))$ with respect to $\mu_{\mathrm{B}}(/\mathtt{b}/)$, the model-mean for a canonical /b/, and $\mu_{\mathrm{S}}([\mathtt{s}])$, the model-mean for a realization [s].

More precisely, since we use three-state HMMs for each phoneme /b/, phone [s] or pair (b: s), the model-mean of the first HMM state represents the acoustics of the initial part of the phoneme, phone or pair, that of the second state the acoustics of the central part, and the model-mean of the third state captures the final part of the phoneme, phone or pair. Any comparison between $\mu_{\mathrm{B}}(/\mathtt{b}/)$, $\mu_{\mathrm{S}}([\mathtt{s}])$ and $\mu_{\mathrm{B:S}}((\mathtt{b}:\mathtt{s}))$ in the following will imply a set of three comparisons, one each between the model-means of the *corresponding states* of the three-state models.

These model-means are 39-dimensional vectors, which makes visualization somewhat challenging. However, since any three points in Euclidean space lie on a 2-dimensional plane, we can plot the relative locations of the three model-means in a standard x–y plane.

Some normalization of inter-point distances is desirable in order to extend this idea from viewing a single triple, e.g., $\{\mu_{\mathrm{B}}(/\mathtt{ae}/), \mu_{\mathrm{B:S}}((\mathtt{ae}:\mathtt{eh})), \mu_{\mathrm{S}}([\mathtt{eh}])\}$, to simultaneously viewing many triples. To this end, we map two points of each triple to two fixed points in the x-y plane and scale the coordinates for the third point such that the three *relative* distances between the three means are preserved. In particular, $\mu_{\mathrm{B}}(/\mathtt{b}/)$ is mapped to the origin (0,0), $\mu_{\mathrm{S}}([\mathtt{s}])$ is mapped to (1,0) on the x-axis, and $\mu_{\mathrm{B:S}}((\mathtt{b}:\mathtt{s}))$ to the positive-y half-plane while preserving relative distances. Details of this procedure are relegated to Appendix A.

Fig. 1 displays the results of projecting the model-means for all triples on to a common 2-dimensional plane in this manner – one triple for each state of every (b: s) pair for which we have sufficient data. The plot in the center shows the location of $\mu_{\mathrm{B:S}}((\mathtt{b}:\mathtt{s}))$, normalized for $\mu_{\mathrm{B}}(/\mathtt{b}/) = (0,0)$ and $\mu_{\mathrm{S}}([\mathtt{s}]) = (1,0)$.

To help interpret the results of the plot in the center, we also replace $\mu_{\mathrm{B:S}}((\mathtt{b}:\mathtt{s}))$ in the triple $\{\mu_{\mathrm{B}}(/\mathtt{b}/), \mu_{\mathrm{B:S}}((\mathtt{b}:\mathtt{s})), \mu_{\mathrm{S}}([\mathtt{s}])\}$ with $\mu_{\mathrm{B:S}}((\mathtt{b}:\mathtt{b}))$ for each state of each pair (b: s), while preserving the scaling factor $|\mu_{\mathrm{S}}([\mathtt{s}]) - \mu_{\mathrm{B}}(/\mathtt{b}/)|$, and obtain the plot to the left in Fig. 1. This plot corresponds to the location of the average acoustics of a /b/ realized as a [b]. Similarly we obtain the location of the average acoustics $\mu_{\mathrm{B:S}}((\mathtt{s},\mathtt{s}))$ of an /s/ realized as an [s]. These are plotted on the right in Fig. 1.

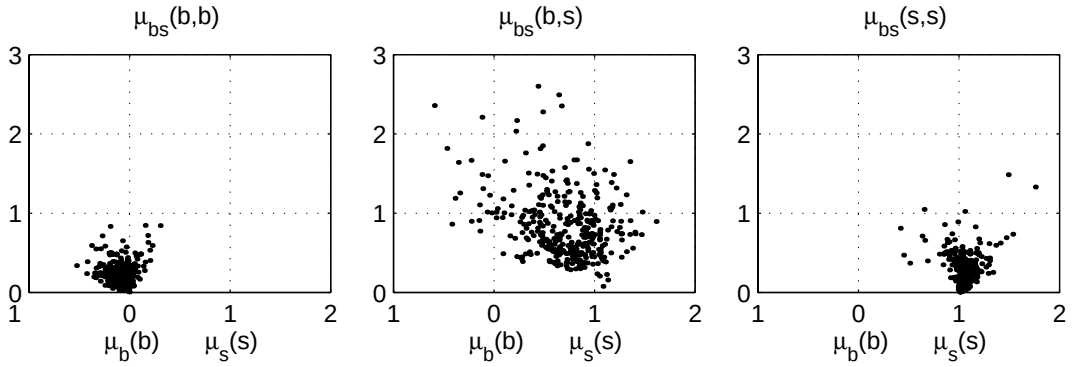

Fig. 1. Comparison of Average Acoustics in the Normalized Acoustic Space. In the normalized space $\mu_{\mathrm{B}}(/\mathtt{b}/)$ is mapped to $(0,0)$, $\mu_{\mathrm{S}}([\mathtt{s}])$ is mapped to $(1,0)$ so that the distance between these two points is unity and relative distances are preserved.

- The plot on the left shows that the acoustics of a /b/ realized as a [b] are all crowded around the model-mean, which is at (0,0), and similarly for an /s/ realized as an [s] around (1,0), as shown by the plot on the right in Fig. 1.
- Compared to these canonical pronunciations, things are much more variable when a pronunciation change occurs. Even when a realization is labeled as an [s] by a human labeler, the acoustics are widely scattered around the model-mean for an [s]. Furthermore, note that the spread is *not isotropic* around [s]: there is a distinct bias in the surface acoustics towards the acoustics of the canonical phoneme /b/. In many instances the observed acoustics when a phoneme /b/ is labeled [2] to have been realized as a phone [s] are actually closer to those of a /b/ realized as a [b] than to those of an /s/ realized as an [s]!

The conclusions supported by these plots are that

1. the acoustics of a phoneme /b/, when realized as a phone [s], lie somewhere between the average realization of the phoneme /b/ and the average realization of the phone [s];
2. neither the phoneme /b/ nor the phone [s] provide a good fit for these realizations; and
3. pronunciation change is *spectrally* partial.

The last conclusion is justified by the nature of the acoustic front-end of the ASR system which, loosely speaking, extracts the speech energy in specific spectral bands, and performs some linear (or monotonic) transformations in order to produce the observations **A**.

2.2. Acoustic likelihood of alternate realizations

It is clear from the previous section that pronunciation change is often partial. A natural question to ask is whether traditional pronunciation modeling techniques are able to accommodate this phenomenon.

[2] We dogmatically avoid designating an instance a *labeling error* even if the acoustics indeed locally resemble those of an /ae/ realized as an [ae], partly because the labelers listen to a wider context while labeling.

In order to understand how best to model the acoustics of a (b: s) pair in an ASR system, we compare the likelihood assigned to the acoustics by the three HMMs $P_{\mathbf{B}}$, $P_{\mathbf{S}}$ and $P_{\mathbf{B:S}}$ discussed above. For each segment of the acoustics, both in the training set and the phonetically labeled 30 min of test data, we have the canonical phonemic transcription $\hat{\mathbf{B}}$, the manual phonetic labels $\hat{\mathbf{S}}$, and their alignment (pair labels $\widehat{\mathbf{BS}}$). The phonetic inventories used for the canonical and manual transcriptions are the same, [3] allowing us to compute likelihoods with four model+transcription combinations:

- $P_{\mathbf{B}}(\mathbf{A}|\hat{\mathbf{B}}) \equiv$ the case when canonical pronunciations are used for both HMM training and test-likelihood computation, and no pronunciation modeling is performed whatsoever;
- $P_{\mathbf{B}}(\mathbf{A}|\hat{\mathbf{S}}) \equiv$ the case when canonical pronunciations are used for HMM training, but manual phonetic labels for test-likelihood computation, providing insight into methods in which no change is made to the conventional acoustic model training procedure, but elaborate new dictionary entries are created for alternate pronunciations and used during recognition;
- $P_{\mathbf{S}}(\mathbf{A}|\hat{\mathbf{S}}) \equiv$ the case when manual phonetic labels are used for both HMM training and test-likelihood computation, providing insight into methods in which a rich set of pronunciations is used both during training and recognition; and
- $P_{\mathbf{B:S}}(\mathbf{A}|\widehat{\mathbf{BS}}) \equiv$ the non-standard case when base-form/surface-form pairs are used for both HMM training and test-likelihood computation.

Concentrating our attention on the subset of acoustic frames that align with locations of pronunciation change, as determined by comparing the canonical phonemic transcription $\hat{\mathbf{B}}$ and the manual phonetic labels $\hat{\mathbf{S}}$, we find the total likelihoods to be ordered as

$$P_{\mathbf{B:S}}(\mathbf{A}|\widehat{\mathbf{BS}}) > P_{\mathbf{S}}(\mathbf{A}|\hat{\mathbf{S}}) > P_{\mathbf{B}}(\mathbf{A}|\hat{\mathbf{S}}) > P_{\mathbf{B}}(\mathbf{A}|\hat{\mathbf{B}}).$$

Fig. 2 summarizes the results of computing these likelihoods, averaged over instances in the training data and, separately, the test data when a base-form phoneme is marked to have been realized as a different phone.

First of all, we see that $P_{\mathbf{B}}(\mathbf{A}|\hat{\mathbf{S}})$, is higher than $P_{\mathbf{B}}(\mathbf{A}|\hat{\mathbf{B}})$, although not by much, which helps to explain the moderate gains obtained when pronunciation models are applied during recognition of spontaneous speech by an ASR system, even when the acoustic models are trained only on canonical pronunciations.

More importantly, $P_{\mathbf{S}}(\mathbf{A}|\hat{\mathbf{S}})$ is considerably higher than $P_{\mathbf{B}}(\mathbf{A}|\hat{\mathbf{B}})$. This shows that using surface-form transcription for acoustic model training is advisable, especially if they are to be subsequently employed in an ASR system with a rich model of alternate pronunciations.

Finally, $P_{\mathbf{B:S}}(\mathbf{A}|\widehat{\mathbf{BS}})$, the likelihood assigned by models based on phone-pairs, is much higher than all the others. If it were only on the training set, this dramatic increase in likelihood could be explained away by the fact that there are many more parameters (HMMs) in $P_{\mathbf{B:S}}$ than either $P_{\mathbf{B}}$ or $P_{\mathbf{S}}$. However, the considerable gain in the test-set likelihood belies this fear of over-fitting. This strongly suggests the investigation of *joint* acoustic models of frequently observed base-form/surface-form pairs, a topic addressed in Section 4.

[3] The symbol set used by the transcribers is very slightly different from the PronLex phone set used in our base-form dictionary. However, since all our acoustic models use the PronLex phone set, we map the actual manual labels to this set. This mapping does not significantly affect any results or conclusions of this paper.

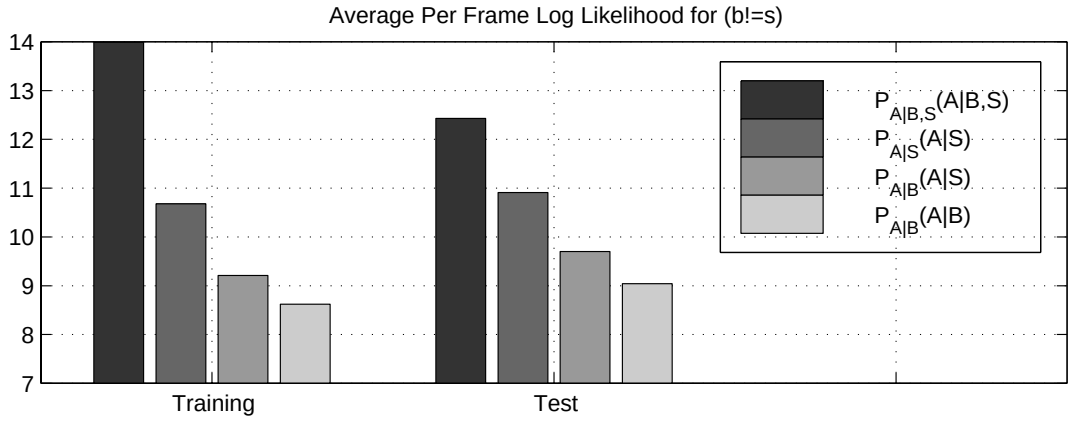

Fig. 2. Average per frame log likelihood of training and test data.

2.3. Temporal characteristics of alternate realizations

So far we have intentionally ignored the fact that the acoustics of a realization of a phone are actually represented by a temporal sequence of feature vectors **A**. In this section, we focus on the sequential or temporal properties of the acoustic realizations during pronunciation change. In order to see why the temporal characteristics are important in terms of partial pronunciation change, consider an example where the word HAD which has the base-form /hh ae dd/ is labeled as [hh ae dd] by one human transcriber, i.e., no pronunciation change is noted, but as [hh eh dd] by another human transcriber, noting a pronunciation change (ae: eh). In this case the surface-form representation of the base-form /ae/ is unclear. Now assume that the first half of the acoustic realization of the base-form /ae/ actually sounds like [ae] but the second half sounds like [eh]. This could explain why the two transcribers disagree. This kind of partial change would be better modeled in ASR if the surface-form representation of the alternate pronunciation of /ae/ permitted a finer temporal resolution than the entire span of the vowel.

Do such realizations exist in real speech?

If they do, how frequent are they?

In order to answer these questions we conduct an experiment using the acoustic models estimated from the base-form transcriptions, $P_{\mathbf{B}}$, together with the canonical pronunciations $\hat{\mathbf{B}}$ and the manual phone labels $\hat{\mathbf{S}}$. Recall that we are using three state left-to-right HMMs and each state can produce one or more consecutive frames of acoustic observations **A**. We investigate pronunciation changes at a temporal granularity finer than the entire phone by adding new transitions to the graph topology of the original HMM structure.

We continue to use the context independent acoustic models with single Gaussian pdfs trained as described above. We again focus only on instances in the test data when a phoneme is substituted with another phone. For each such instance we investigate whether allowing partial pronunciation change from the canonical to the labeled alternate realization at a subphonetic level, rather than only allowing a change in the entire phonetic realization, would result in higher likelihood. To accomplish this, two additional HMMs are constructed from a combination of states of the base-form and surface-form trained HMMs $P_{\mathbf{B}}$ and $P_{\mathbf{S}}$, respectively, as illustrated in Fig. 3.

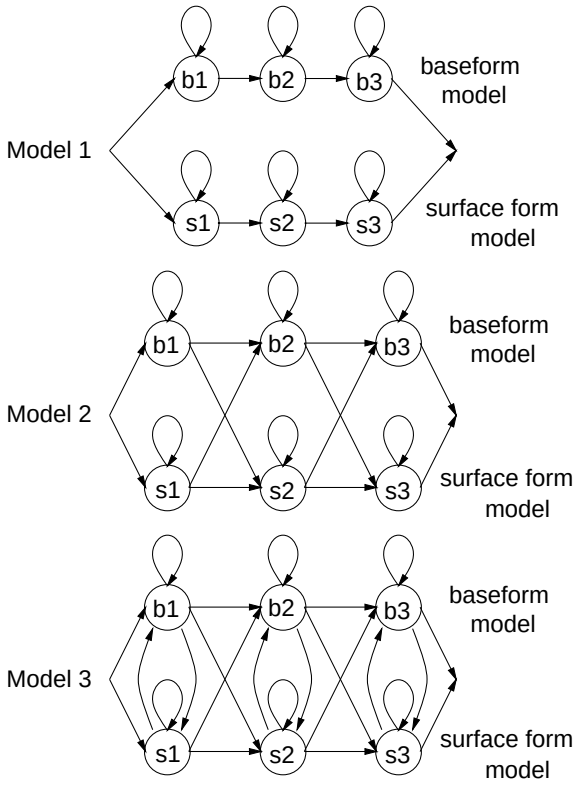

Fig. 3. Models used for increasing the temporal resolution.

Note that the first model in Fig. 3 allows only paths corresponding to a change in pronunciation of the entire phoneme, the second allows a change in the pronunciation of any of the initial, central or final segments of a phoneme, and the third allows every 10 ms acoustic frame in **A** to be "pronounced" in one of the two ways, respecting only the sequential ordering of the three subphonetic segments. We align the acoustic realization of a phone against each of these three models and compute the most likely HMM state sequence for the realization. The exact HMM transition probabilities, which are not available for the newly added transitions but which do not contribute significantly to the acoustic likelihood in any event, are ignored in this likelihood calculation, and it is instead assumed that all the permissible transitions have equal probabilities.

By construction, all models permit paths corresponding to a pronunciation change in the entire phonetic segment; this corresponds to the null hypothesis, as manually labeled, that the pronunciation of the entire phoneme has been changed. Having obtained the most likely path, we then count the number of times (i.e., the number of test tokens for which) this best path consists of a sequence of HMM states corresponding entirely to the manually marked surface-form ($s_1s_2s_3$), entirely to the base-form ($b_1b_2b_3$), and to some temporal *patchwork* of the two, e.g., $s_1b_2b_3$ or $b_1s_2b_3$. The relative frequency of the three outcomes, computed over all instances in the test corpus that were manually labeled as containing a phone-substitution, is reported in Table 1.

The results of Table 1 indicate that

Table 1
Distribution of the best state sequence when a canonical phoneme /b/ is realized as a surface phone [s]

Model	Best path		
	$b_1b_2b_3$	Others	$s_1s_2s_3$
Model 1	41%	–	59%
Model 2	20%	50%	30%
Model 3	15%	62%	23%

1. *the pronunciation change is indeed partial*, since even when a change from /b/ to [s] is marked by a labeler, a significant minority (41%) of tokens align better with the model for /b/ when $b_1b_2b_3$ and $s_1s_2s_3$ are the only two possible paths;
2. in a large majority of instances of pronunciation change (50–62%), the highest-likelihood path corresponds to a temporal switching between the base-form and the surface-form.

Therefore pronunciation change is temporally partial and increasing the temporal resolution of the acoustic–phonetic model may be needed to effectively model partial changes in an ASR system. Section 4 discusses such a pronunciation model.

Note from the last row of Table 1 that 15% of the tokens marked by human labelers to have undergone a change from their canonical pronunciation /b/ to a realization [s] actually align better with the model for /b/ than that for [s], providing further evidence of partial pronunciation change. Statistics of interlabeler agreement, discussed in Section 3, will further support this argument.

A possible criticism of the experiment of Table 1 is that the two HMM topologies corresponding to temporally partial changes in pronunciation, shown in Fig. 3, simply admit additional state sequences not permitted by the HMM topology at the top of the figure. One may suspect that the statistics in Table 1, where the most likely path is something other than $b_1b_2b_3$ or $s_1s_2s_3$, are merely a matter of chance given the additional degrees of freedom in the alternate HMM topologies. We attempt to alleviate this concern, and reassert that the pronunciation change is indeed temporally partial, by conducting the following two experiments.

1. We reuse the model topology of the third HMM (Model 3) of Fig. 3, but instead of choosing $s_1s_2s_3$ to be the HMM states of the marked alternate pronunciation [s] of the phoneme /b/, whose HMM states are $b_1b_2b_3$, we randomly pick another phone, say [z], and choose its HMM states $z_1z_2z_3$ to construct Model 3 of Fig. 3. We align the acoustics against this composite HMM and collect the same statistics as the third line of Table 1, namely, the fraction of tokens in which the best path is made up entirely of $b_1b_2b_3$ or $z_1z_2z_3$ and the fraction of tokens in which it corresponds to some temporal patchwork of /b/ and [z]. We repeat this exercise, each time with a different phone in the role of [z], *average* the statistics thus obtained, and present them on the second row of Table 2.
2. We repeat the experiment above, but now we keep [s] fixed and replace /b/ in turn with each permissible phoneme, say /p/, and collect statistics on the fraction of times the best path is not one of $p_1p_2p_3$ or $s_1s_2s_3$. These figures are presented on the third row of Table 2.

It is clear from Table 2 that the fraction of state sequences consisting of a temporal patchwork of base-form and surface-form states is much higher (62%) when the HMM is made up of the *correct* base-form and surface-form combination, namely /b/ and [s], than for another HMM

Table 2
Average distribution of the best state sequence for Model 3 when a phoneme /b/ is realized as a phone [s]

Model 3, Base-form states	Model 3, Surface-form states	$b_1 b_2 b_3$ or $p_1 p_2 p_3$	Others	$s_1 s_2 s_3$ or $z_1 z_2 z_3$
$b_1 b_2 b_3$	$s_1 s_2 s_3$	15%	62%	23%
$b_1 b_2 b_3$	$z_1 z_2 z_3$	58%	37%	5%
$p_1 p_2 p_3$	$s_1 s_2 s_3$	3%	32%	65%

/p/ denotes any other phoneme and [z] any other phone which replaces /b/ and [s], respectively, in the model.

with the same topological freedom but with a *random* alternate phone – [z] instead of [s] or /p/ instead of /b/.

3. Automatic phonetic transcription of acoustic data

Modern speech recognition systems are often trained with tens if not hundreds of hours of speech from a multitude of speakers recorded under varying acoustic conditions. Producing the orthographic transcriptions of the speech and a dictionary of canonical pronunciations to cover all the spoken words already entails considerable cost. While the results of the preceding section clearly demonstrate the greater accuracy of acoustic models of base-form/surface-form pairs compared to models of base-forms alone, an obvious limitation in putting this idea into practice is the unacceptably high cost, not to mention the delay in the system development cycle, of manually producing phonetic transcriptions of tens or hundreds of hours of spontaneous speech. Thus it is natural to seek means for automatic phonetic transcription.

Several methods have been proposed (cf., e.g., Saraçlar et al., 2000) for automatically generating a large corpus of phonetically transcribed speech, particularly when the orthographic transcriptions of the speech, a dictionary of canonical pronunciations, and a modest amount of phonetically transcribed speech are available. The usual bootstrapping procedure is to

1. train context-independent acoustic–phonetic models $P_{\mathbf{S}}$ based on phonetic transcriptions from the (small) manually transcribed portion of the speech corpus;
2. align the canonical and the actual pronunciations in the phonetically transcribed portion, get $\widehat{\mathbf{BS}}$, and obtain a *pronunciation model* that captures all frequent phoneme-to-phone changes;
3. apply the pronunciation model of the previous step to the canonical pronunciations in the large corpus, producing a network of locally alternate pronunciations for all words in all the utterances in the large corpus;
4. use the previously obtained acoustic models to align the speech with this phonetic-network and extract the most likely phonetic (surface-form) transcription of the utterances.

This entire 4-step procedure of training acoustic models on the surface-form transcriptions, building the pronunciation model, creating pronunciation networks and retranscribing the speech may then be iterated on the large speech corpus in the hope of incremental improvement.

To test the effectiveness of this procedure, we trained context-independent HMMs with a 4-Gaussian mixture density for each HMM state on the phonetically transcribed portion of our training set. We also estimated a decision tree pronunciation model, as described by Saraçlar et al., 2000, and used these two to phonetically transcribe the 30-min portion of our test set with

manual phonetic transcriptions. The phone error rate (PER) of the automatic transcription relative to the manual transcription, including substitution-, deletion- and insertion-errors, was 26.6%.

It seems natural to ask, particularly given the partial nature of pronunciation change demonstrated in the preceding section, if the accuracy of this transcription can be improved any further. Furthermore, since this accuracy is being measured against a gold standard provided by a human labeler, it seems natural to ask about the rate of interlabeler agreement on this task. As it happens, a small set of utterances (~2000 phones) in the phonetically labeled corpus has been independently labeled by two human transcribers, and we use this portion in an attempt to answer these two questions.

Interlabeler agreement of "ca. 75–80%" has been reported on this task Greenberg (1998). Since this is quite comparable to the agreement between our automatic phonetic transcription and the manual transcription (26.6% PER), we undertook a 3-way supervised alignment between our automatic transcription and the two manual transcriptions of the 2000-phone subset. An actual example of this alignment for the word PARENTS, and the overall proportion of each type of (dis)agreement in the 2000-phone subset is given in Table 3.

As these results indicate, the automatic transcriptions fare almost as well as the human transcribers in terms of overall PER. If the gold standard is taken to be the transcriptions produced by the first transcriber (T1) the PER of the automatic transcriptions on this (albeit small) set is 25.7% whereas the mismatch between the two human transcribers is 24.7%.

If one focuses on the (75.3%) cases where the human transcribers T1 and T2 agree, the PER of automatic transcription is still 14.5%, which shows that there is some room for further improvement of the automatic transcription process described here. It shows that *errors* in the automatic phonetic transcription, no matter which human transcriber is used as the gold standard, are not confined to the (24.7%) instances in which there is disagreement between the human transcribers. More remarkably, however, the PER between the automatic and either human transcription jumps to over 60% in the regions of pronunciation ambiguity, i.e., in the instances where the two human transcribers disagree.

Table 3
Example alignment segment and proportions of agreement of each type among the two human labelers (Tl and T2) and the automatic transcription (A)

Labeler			Agreement Type			Overall proportion
T1	T2	A	T1 vs T2	A vs T1	A vs T2	
p	p	p	✓	✓	✓	64.4%
eh	ae	ae			✓	8.0%
r	r	r	✓	✓	✓	
ax	–	ih				6.4%
n	en	n		✓		10.3%
t	t	–	✓			10.9%
S	S	S	✓	✓	✓	
Total agreement			75.3%	74.3%	72.4%	

From a modeling point of view, the genuine ambiguity about the correct transcription in the latter cases is a good motivation for modeling the surface-form representation as a latent variable, instead of seeking its explicit realization, during HMM parameter estimation and speech recognition.

4. Speech recognition experiments

Two sets of speech recognition experiments have been conducted to evaluate the performance of the acoustic models designed to handle partial pronunciation change of the sort described in Section 2 above. The Switchboard corpus is used in our experiments. A vocabulary of approximately 20,000 words provides adequate coverage for the corpus. We use 60 h 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 2 h of speech (19 entire conversations, 2427 utterances, 18,100 words) is set aside, as mentioned earlier, for testing the ASR system. A 30-min subset of the test set (451 utterances, 18,000 phones) is phonetically labeled.

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, which has a single pronunciation for approximately 94% of the words in the test vocabulary, two pronunciations for a little over 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.

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

Without *any* pronunciation modeling, the ASR output following trigram rescoring has a WER of 39.4%. The 30-min subset, for which manual phonetic transcriptions are available, has a WER of 49.1%. These are our baseline figures for further comparisons.

We could not find a systematic cause for the WER on the 30-min subset being significantly higher than the remainder of the test set. Utterances for phonetic transcription were selected a priori by Greenberg (1998), using criteria not known to us, and this resulted in many conversations containing a few phonetically labeled utterances each, including the conversations in our test set. No further characterization of these utterances that may explain the difference in WER is obvious.

4.1. Explicit pronunciation models

In the first set of experiments, we use as a pronunciation model an explicit listing of the canonical and alternate pronunciations of words in the recognition dictionary (see Riley et al.,

1999). We compare via lattice rescoring the three acoustic models described in Section 2: $P_{\mathbf{B}}$ and $P_{\mathbf{S}}$, which differ in the transcriptions on which they were trained but which use the same phonetic inventory, and $P_{\mathbf{B:S}}$, which is trained on the pair transcriptions.

Acoustic model training using the pair transcriptions is not entirely straightforward, especially since we are using decision tree based state clustered triphones. We build individual decision trees for each surface-form phone, allowing questions about its own base-form phoneme and its context during clustering. In order to limit complexity, we only include the base-form phoneme identity of the neighboring pairs in the context. We explored many alternative strategies – building individual decision trees for each base-form phoneme as well as building a single decision tree – but found them to be less effective. Interested readers are referred to the doctoral dissertation of Saraçlar (2000) for details.

Incorporating alternate pronunciations during rescoring: For example, if the word `HAD` is hypothesized on a link in the lattice, and `HAD` happens to have two pronunciations `/hh ae dd/` and `[hh eh dd]` in the dictionary, the former being the canonical pronunciation, then

- in case of $P_{\mathbf{B}}$ and $P_{\mathbf{S}}$, we replace the word-link for `HAD` with two parallel paths, with the phone sequence `hh ae dd` on one path and `hh eh dd` on the other, and appropriate triphone HMMs for the four phone(me)s – `hh`, `ae`, `eh` and `dd` – are used to compute acoustic scores on the two paths, whereas
- in case of $P_{\mathbf{B:S}}$, the symbol sequences on the two paths are (`hh:hh`) (`ae:ae`) (`dd:dd`) and (`hh:hh`) (`ae:eh`) (`dd:dd`), and appropriate triphone HMMs for the four base-form/surface-form pairs – (`hh:hh`), (`ae:ae`), (`ae:eh`) and (`dd:dd`) – are used to calculate acoustic scores.

Such operation are carried out on all lattice links by finite-state operations and the resulting phone-graph is rescored using the appropriate acoustic models.

The word error rate (WER) measured against the word level transcriptions for the phonetically annotated 30 min portion of the test set are presented in Table 4. For illustration, we also present in the same table the phone error rate (PER) of the same ASR output measured against either (i) the base-form transcription $\hat{\mathbf{B}}$ or (ii) the hand-labeled surface-form transcription $\hat{\mathbf{S}}$ as the gold standard.

1. Note that on the first line of Table 4 the PER of the ASR output measured against $\hat{\mathbf{B}}$ is much lower than the PER of the same output measured against $\hat{\mathbf{S}}$ – 34.7% vs 48.1%. We conjecture that this is because the acoustic models $P_{\mathbf{B}}$ are trained on the canonical transcriptions $\mathbf{B}$.
2. What is perhaps quite surprising is that the models $P_{\mathbf{S}}$ trained on the more phonetic transcriptions $\hat{\mathbf{S}}$ result in a higher word error rate, while considerably lowering the PER relative to the manual transcription $\hat{\mathbf{S}}$ – from 48.1% to 43.6%. This is perhaps because of the increased confusability caused by adding alternate pronunciations to the dictionary. Table 5 illustrates how the word `HAD` may now be confusable with `HEAD`, and error reduction from more accurate recognition of the phonemes may yet be undone by their incorrect mapping to words.

Table 4
Recognition performance of acoustic models with a rich pronunciation dictionary

Acoustic model	PER wrt $\hat{\mathbf{B}}$	PER wrt $\hat{\mathbf{S}}$	WER
$P_{\mathbf{B}}$	34.7%	48.1%	49.0%
$P_{\mathbf{S}}$	42.9%	43.6%	50.6%
$P_{\mathbf{B:S}}$	33.8%	43.9%	47.8%

Table 5
Dictionary for **S** has an increased number of homophones

Word exemplars	Dictionary Pronunciations		
	B	**S**	**B : S**
HAD	hh ae dd	hh ae dd	hh: hh ae: ae dd: dd
HAD	–	hh eh dd	hh: hh ae: eh dd: dd
HEAD	hh eh dd	hh eh dd	hh: hh eh: eh dd: dd

3. We note that acoustic models trained on the pair-transcription $\widehat{\mathbf{BS}}$ achieve the ideal compromise – the PER of the base-form-side of the ASR output with respect to the canonical transcription is lower than that for $P_{\mathbf{B}}$, the PER of the surface-form-side of the ASR output is about as good as that for $P_{\mathbf{S}}$, and the WER is lower than that of either of the other two models. In this case, the increased phonetic accuracy of the acoustic models is not undone by increased lexical confusability, as shown by the sample dictionary entries in Table 5.

4.2. Implicit pronunciation models

In the second set of experiments, we use a more recent method for pronunciation modeling called *state level pronunciation model* or SLPM (Saraçlar et al., 2000), which accommodates alternate surface-form realizations of a phoneme by allowing the HMM states of the model of a base-form phoneme /b/ to share output densities with models of its alternate surface-form realizations [s]. This "merging" of the two acoustic models therefore captures both spectrally partial changes and temporally subphonetic changes in pronunciation. For the sake of completeness, we briefly summarize the SLPM construction procedure from Saraçlar et al. (2000).

4.2.1. State level pronunciation models

To recapitulate the mechanism of state level pronunciation modeling via an example, we suppose that the word HAD has a canonical pronunciation /hh ae dd/ but it may sometimes be realized as [hh eh dd], which happens to be the canonical pronunciation of the word HEAD. The sketch at the top of Fig. 4 illustrates how the two alternate pronunciations for HAD will be represented at the phone level, and the sketch in the middle shows how a context-independent HMM system utilizing *explicit pronunciation modeling* could permit these alternatives in a recognition network.

The SLPM instead performs *implicit pronunciation modeling* as illustrated by the sketch at the bottom of Fig. 4. 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.

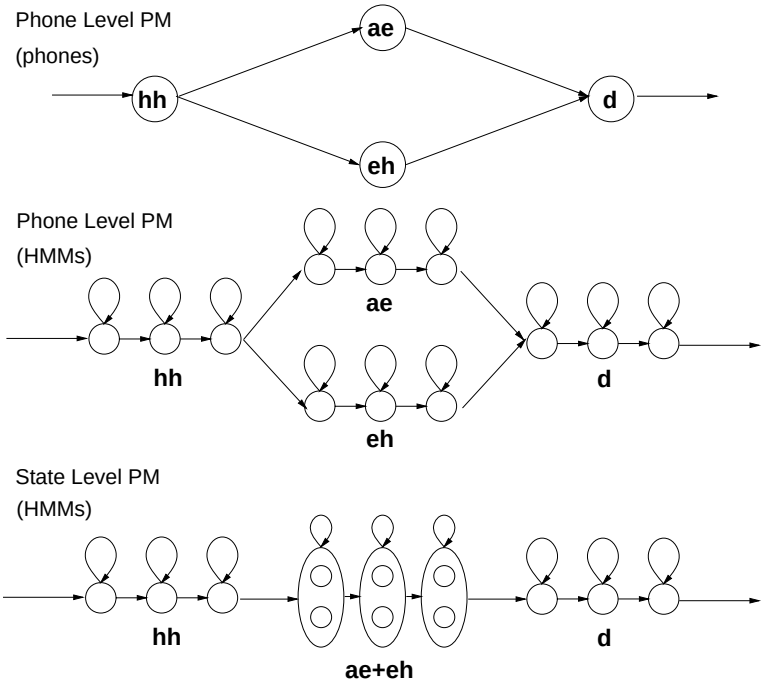

Fig. 4. The effect of allowing phoneme /ae/ to be realized as phone [eh], context independent models.

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 (Young et al., 1995). A separate clustering tree is grown for each of the three states in an HMM. The detailed effect of the SLPM on such a system is illustrated in Fig. 5. 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 Fig. 5. 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* of the phoneme /d/. This is illustrated by the sketch on the right in Fig. 5.

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 an HMM state is illustrated in Fig. 6.

The HMMs used for modeling the alternate realization in the SLPM may either be the same as the set of baseline HMMs $P_{\mathbf{B}}$, or a different set of HMMs $P_{\mathbf{S}}$ trained on the surface-form transcriptions $\hat{\mathbf{S}}$. Both alternatives have been investigated by Saraçlar et al. (2000), with the aid of an automatic boot-strapping procedure for generating the surface-form transcriptions $\hat{\mathbf{S}}$ for the entire 60-h acoustic training corpus.

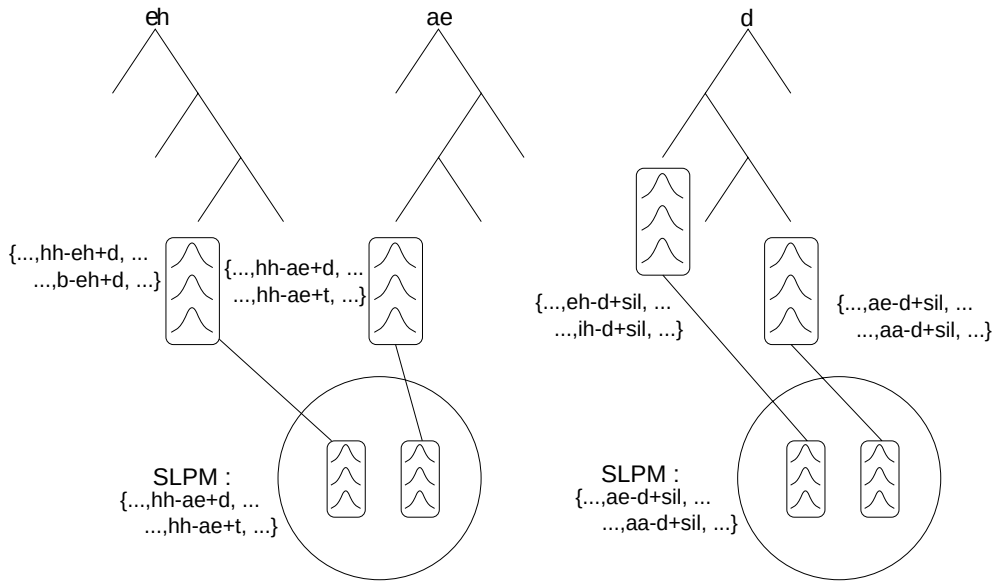

Fig. 5. Sharing Gaussian mixtures among different HMM states to accommodate the pronunciation change /ae/ →[eh].

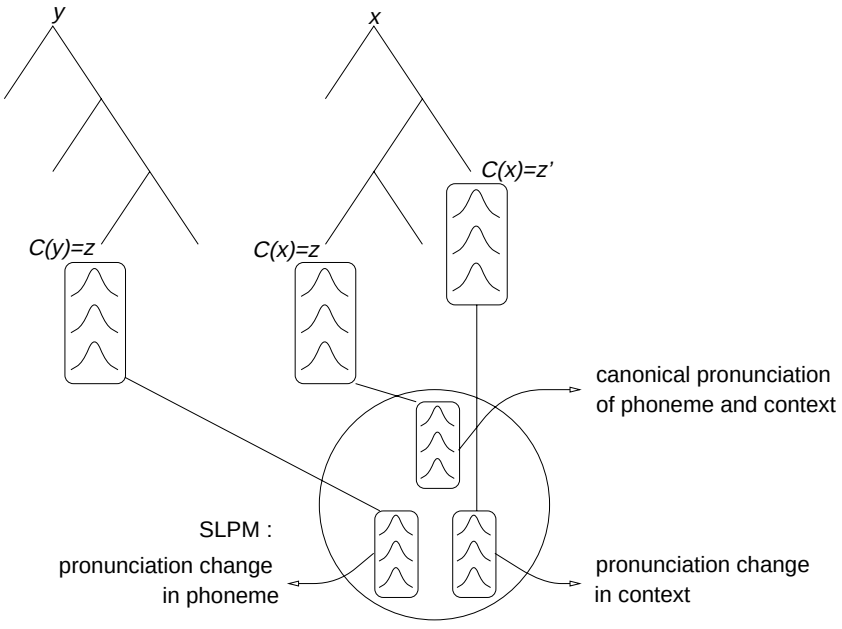

Fig. 6. Overall effect of sharing Gaussian mixtures among different tree-clustered HMM states to accommodate pronunciation changes $x \rightarrow y$ and $z \rightarrow z'$ where z and z' are "contexts" of x.

Table 6
Performance of the SLPM on the phonetically annotated subset and the entire test set

Acoustic models combined during SLPM construction	PER wrt $\hat{\mathbf{B}}$	PER wrt $\hat{\mathbf{S}}$	WER (Subset)	WER (Full)
Riley et al. (1999) (Baseline)	–	–	–	38.9%
$P_{\mathbf{B}}$ with itself	34.6%	50.1%	49.0%	39.0%
$P_{\mathbf{B}}$ and $P_{\mathbf{S}}$	33.1%	48.5%	47.9%	38.2%
$P_{\mathbf{B}}$ and $P_{\mathbf{B:S}}$	32.8%	49.4%	47.2%	37.7%

Key insight. It was shown by Saraçlar et al. (2000) that acoustic models $P_{\mathbf{B}}$ estimated from the base-form transcriptions can be effectively combined with acoustic models $P_{\mathbf{S}}$ estimated from the surface-form transcriptions in the SLPM framework. We hope the reader is convinced following the analysis of Sections 2 and 3 that the considerable success of their method is due to the fact that the SLPM is able to model partial pronunciation change. We further believe that the same phenomenon is also being modeled by the pronunciation modeling technique proposed by Hain and Woodland (1999). We call them implicit pronunciation modeling techniques for obvious reasons.

A natural question that follows is whether the SLPM, which combines $P_{\mathbf{B}}$ and $P_{\mathbf{S}}$, could be improved upon by using $P_{\mathbf{B:S}}$. This is the subject of the following.

4.2.2. Additional ASR experiments with the SLPM

Here, we contrast merging the baseline models $P_{\mathbf{B}}$ with the surface-form trained models $P_{\mathbf{S}}$, which was described in Saraçlar et al. (2000), against the alternative of merging $P_{\mathbf{B}}$ with the models $P_{\mathbf{B:S}}$, which were shown in Section 2 to better model the acoustics of partial pronunciation change. Table 6 shows the results of these experiments where, in addition to the phonetically labeled 30 min portion, WER is also reported on the entire 2-h test set.

Under identical training and test conditions, Riley et al. (1999) report that techniques which account for complete pronunciation change but not partial pronunciation change improve the overall WER from 39.4% to 38.9%, and this may be considered an alternate baseline for our experiments. Accounting further for the partial change using our techniques leads to more significant improvements, as seen on the last two lines of the table.

In particular, augmenting the Gaussian densities of the HMM states of, say, `/ae/` in $P_{\mathbf{B}}$ with Gaussian densities of the HMM states of `[eh]` in $P_{\mathbf{S}}$ enables the new HMM to cover realizations of `/ae/` which undergo only a partial change of pronunciation (cf. middle-box of Fig. 1). Alternately, augmenting the mixture for `/ae/` in $P_{\mathbf{B}}$ with Gaussian densities of the HMM states of (`ae: eh`) in $P_{\mathbf{B:S}}$ is even better, since the latter captures not just the surface realizations of `[eh]`, but specifically of `[eh]` whose canonical pronunciation would have been `/ae/`.

The overall reduction in WER is statistically significant for both pronunciation modeling techniques: WER reduction from 38.9% to 38.2% is significant at a p-value of 0.003 and to 37.7% at a p-value < 0.001. The difference between 38.2% and 37.7% is less significant, with a p-value of 0.07. All significance tests were performed using standard methods described by Pallett et al. (1990) and with software provided by the US National Institutes of Standards and Technology.

5. Concluding remarks

In summary, here are the main points of this article.

- We have presented acoustic evidence which demonstrates the prevalence of spectrally and temporally partial pronunciation change in spontaneous conversational speech. We have shown how these partial changes make the very notion of phonetic transcription, be it manual or automatic, a difficult one to define.
- We have presented means for automatically generating phonetic transcriptions whose accuracy rivals interlabeler agreement, and hence may be considered almost as good as human phonetic labeling.
- We have shown a method for using such automatically generated phonetic transcriptions of the acoustic training data to train models which improve speech recognition accuracy by accommodating pronunciation ambiguity. A 1.7% (absolute) WER improvement over a system which does not employ any pronunciation modeling is demonstrated on Switchboard (from 39.4% baseline WER to 37.7%).

The improvement from the last method is greater than that of most previously known methods for pronunciation modeling. The improvements in Hain (2002) are comparable to ours and on a comparable test set.

Finally, our analysis provides an understanding of the reasons why implicit pronunciation modeling techniques, which include the SLPM and those investigated by Hain (2002) and others, are more successful than explicit modeling techniques for conversational speech. Kam et al. (2003) have recently developed further extensions to the SLPM by using HMM adaptation techniques.

Appendix A. Mapping 39-dimensional means to 2-dimensional space

In this appendix, we define the procedure for mapping three points x, y, z in n-dimensional space to points x', y', z' in 2-dimensional Euclidean space, while preserving the relative distances between the three points. We also want x' and y' to be mapped to fixed points in the plane, so we assume $x \neq y$.

Let $D(\cdot,\cdot)$ denote the distance measure in the original n-dimensional space, and $d(\cdot,\cdot)$ denote the distance measure in the target 2-dimensional space. We define $d(\cdot,\cdot)$ to be the Euclidean distance and we will define $D(\cdot,\cdot)$ later. For non-trivial cases [4] preserving the relative distances requires

$$\frac{D(x,z)}{D(x,y)} = \frac{d(x',z')}{d(x',y')}, \qquad \frac{D(y,z)}{D(x,y)} = \frac{d(y',z')}{d(x',y')}, \qquad \frac{D(y,z)}{D(x,z)} = \frac{d(y',z')}{d(x',z')}.$$

Now let $x' = (0,0)$ and $y' = (1,0)$. This gives $d(x',y') = 1$, and we only need to satisfy

$$d((0,0),z') = \frac{D(x,z)}{D(x,y)}, \qquad d((1,0),z') = \frac{D(y,z)}{D(x,y)}.$$

[4] Assuming $z \neq x$ and $z \neq y$. Otherwise, if $z = x$ then $z' = x'$ and if $z = y$ then $z' = y'$.

The general procedure is as follows:
1. Calculate the pairwise distances $D(x,y), D(x,z), D(y,z)$.
2. Let $x' = (0,0)$ and $y' = (1,0)$.
3. Solve the following equations to obtain $z' = (z_1, z_2)$:

$$z_1^2 + z_2^2 = \left[\frac{D(x,z)}{D(x,y)}\right]^2,$$
$$(1-z_1)^2 + z_2^2 = \left[\frac{D(y,z)}{D(x,y)}\right]^2.$$

The solution is given by:

$$z_1 = \frac{1}{2}\left(1 + \left[\frac{D(x,z)}{D(x,y)}\right]^2 - \left[\frac{D(y,z)}{D(x,y)}\right]^2\right),$$
$$z_2 = \sqrt{\left[\frac{D(x,z)}{D(x,y)}\right]^2 - z_1^2}.$$

Note that this procedure does not depend on the actual definition of the distance $D(\cdot,\cdot)$. In the experiment described in Section 2.1 the points actually represent means of 39-dimensional Gaussian distributions with parameters (μ, σ). In an attempt to use the variances for scale normalization, we define $s_i = \sigma_i^{(x)}\sigma_i^{(y)}$ where $\sigma^{(\cdot)}$ is the variance of the Gaussian associated with each point and use s_i to scale each dimension. The distances $D(\cdot,\cdot)$ are thus defined as:

$$(x,y) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - y_i)^2}{s_i}},$$
$$D(x,z) = \sqrt{\sum_{i=1}^{n} \frac{(x_i - z_i)^2}{s_i}},$$
$$D(y,z) = \sqrt{\sum_{i=1}^{n} \frac{(y_i - z_i)^2}{s_i}}.$$

References

Finke, M., Waibel, A., 1997. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In: Proceedings of the European Conference on Speech Communication and Technology. Rhodes, Greece, pp. 2379–2382.

Fosler-Lussier, E., Williams G., 1999. Not just what, but also when: Guided automatic Pronunciation Modeling for Broadcast News. In: DARPA Broadcast News Workshop. Herndon, Virginia, pp. 171–174.

Godfrey, J., Holliman, E.C., McDaniel, J., 1992. Switchboard: telephone speech corpus for research and development. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing. San Francisco, CA, pp. 517–520.

Greenberg, S., 1998. Speaking in Shorthand – A syllable centric perspective for understanding pronunciation variation. In: Proceedings of the ESCA Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition. Kekrade, Netherlands.

Hain, T., Woodland, P., 1999. Dynamic HMM selection for continuous speech recognition. In: Proceedings of the European Conference on Speech Communication and Technology. Budapest, Hungary, pp. 1327–1330.

Hain, T., 2002. Implicit Pronunciation Modeling in ASR. In: Proceedings of the ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexical Adaptation for Spoken Language. Estes Park, CO.

Jurafsky, D. et al., 2001. What kind of pronunciation variation is hard for triphones to model? In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing. Salt Lake City, UT, pp. 577–580.

Kam, P., Lee, T., Soong, F., 2003. Modeling Cantonese pronunciation variation by acoustic model refinement. In: Proceedings of the European Conference on Speech Communication and Technology. Geneva, Switzerland, pp. 1477–1480.

Ma, K., Zavaliagkos, G., Iyer, R., 1998. Pronunciation modeling for large vocabulary conversational speech recognition. In: Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, pp. 2455–2458.

Mohri, M., Pereira, F., Riley, M., 2000. The design principles of a weighted finite state transducer library. Theoretical Computer Science 231, 17–32.

Pallett, D., Fisher, W., Fiscus, J., 1990. Tools for the analysis of benchmark speech recognition tests. In: Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing. Albuquerque, NM, pp. 97–100.

Riley, M. et al., 1999. Stochastic pronunciation modeling from hand-labeled phonetic corpora. Speech Communication 29, 209–224.

Saraçlar, M., 2000. Pronunciation modeling for conversational speech recognition. PhD Thesis, Johns Hopkins University.

Saraçlar, M., Nock, H., Khudanpur, S., 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language 14, 137–160.

Stolcke, A. et al., 2000. The SRI March 2000 Hub-5 Conversational Speech Transcription System. In: Proceedings of the Speech Transcription Workshop. College Park, MD.

Young, S. et al., 1995. The HTK Book (version 2.0). Entropic Cambridge Research Laboratory, Cambridge, UK.