Properties of Pronunciation Change in Conversational Speech Recognition

Murat Saraçlar and Sanjeev Khudanpur

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD 21218-2686
{murat,sanjeev}@clsp.jhu.edu

ABSTRACT

It is widely acknowledged that pronunciations in spontaneous speech differ significantly from citation form. For this reason, pronunciation modeling has received considerable attention in recent automatic speech recognition literature. Most of the attention however has focussed on describing an alternate pronunciation as a different sequence of phonetic units using the same inventory of phones which describe canonical pronunciations. This approach assumes that the deviation from the canonical pronunciation can be represented in terms of complete changes from the base-form phonemes to the surface-form phones. In this paper, we investigate the properties of pronunciation change in conversational speech and demonstrate that most of the time the change is only partial; 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 environment. The evidence for this behavior comes from the analysis of the acoustic signal, the base-form transcriptions given by the dictionary and the surface form transcriptions given by human labelers. We show that when a pronunciation change occurs, it is often the case that neither the canonical nor the alternate phone represent the acoustics very well. We also show how partial change makes the notion of phonetic transcription, be it manual or automatic, a difficult one. Analysis of manual phonetic transcription of conversational speech reveals a large number (>20%) of cases of genuine ambiguity: instances where human labelers disagree on the identity of the surface form. Based on this analysis, two methods for accommodating partial pronunciation change are developed. The first method attempts to resolve the ambiguity by separately modeling each baseform/surface-form pair. The second method treats the surface form as a hidden variable and "averages out" the ambiguity.

1. Introduction

Acoustic modeling based on phonetic units relies on having an accurate phonetic representation of words. However, the high degree and continuous nature of pronunciation change encountered in conversational spontaneous speech makes such a representation impossible if one insists on having a constant and limited phonetic inventory. An analysis of a portion of the Switchboard corpus labeled by linguists at the phonetic level reveals that the disagreement between human labelers is quite high. This suggests that pronunciation change sometimes yields ambiguous representations when projected onto a limited phonetic inventory. In this paper, we investigate and characterize the properties of pronunciation change in the context of this ambiguity. We differentiate between two types of pronunciation change : *complete change* where the surface form can be clearly identified by humans and *partial change* where even human transcribers cannot agree on the identity of the surface form.

Most of the work on pronunciation modeling tries to predict changes in pronunciation so that words are allowed to have alternate phonetic representations. This sort of explicit pronunciation modeling combined with context dependent acoustic modeling can only partially account for the pronunciation variation in conversational speech as suggested by moderate gains in word error rate reported by various researchers. As opposed to this "linear phonology" approach, "nonlinear" or autosegmental phonological models allow for representations based on asynchronous features and are not constrained by the phonetic inventory. One such model [2, 3] given by Deng is a feature-based phonological model that yields a feature overlapping pattern instead of a phonetic representation. Finke [4] recently proposed using "attribute instances" which include articulatory features, stress etc. as acoustic modeling units and a pronunciation model that predicts variation of these instances. This instance based representation provides a tighter coupling of the pronunciation model and the acoustic model.

In this paper we analyze the properties of pronunciation change to explain the intrinsic ambiguity of phone level transcriptions and we propose methods within the "linear phonology" framework to overcome the problems caused by partial changes. An analysis of the relationship between acoustics and phonemic/phonetic representations is used to explain the recognition results of two methods for improving acoustic modeling using pronunciation modeling. One method extends the units used in acoustic modeling to baseform/surface-form pairs, attempting to resolve the ambiguity by enlarging the inventory and taking a step towards a "tighter coupling" between the acoustic models and the pronunciation model. This approach allows for modeling the acoustics of pronunciation change without increasing lexical confusion, or homophony. Another method models the pronunciation change at the state level so that partial pronunciation changes can be covered. This method also provides more accurate acoustic probabilities for the baseform by keeping the surface form as a hidden variable and summing over all

alternate pronunciations of a baseform. This approach handles ambiguity by averaging instead of disambiguating.

This paper is organized as follows. In Section 2 we present acoustic evidence for partial pronunciation change in conversational speech. The effects of this partial change are further quantified in the interlabeler agreement statistics and in our efforts to obtain accurate phonetic transcriptions by automatic means, as discussed in Section 3 . Finally, speech recognition experiments which accommodate partial pronunciation changes are presented in Section 4 and some concluding remarks are given in Section 5 .

2. Acoustic Analysis of Pronunciation Change

We use the Switchboard corpus, a collection of casual telephone conversations between American English speakers, to study pronunciation changes in conversational speech. A portion (∼4 hours, ∼100K phones) of Switchboard has been phonetically labeled, and it is on this portion of the corpus that our investigations are based. Furthermore, about 30 minutes of this labeled data is in the "test" portion of the corpus, while a little over 3 hours is in the acoustic training set. Model estimation in this section is on the training portion of the hand labeled corpus and evaluation is done on the test portion where appropriate.

In order to understand the nature of pronunciation change and to discover ways of modeling it, we need to investigate the relationship between acoustics and baseform/ surface-form representations.

Consider an occurrence of the word AND which has the baseform /ae n d/ and is labeled as the surface form [eh n d]. In this example, /ae/ is realized as [eh] [1] forming the baseform/surface-form pair (ae,eh). What do the acoustics of this pair look like? If the acoustics are sufficiently similar to those of an [eh], this can be considered as *complete change*, otherwise this is a case of *partial change*. In any case, how should this pair be modeled? Complete pronunciation change may perhaps be dealt with by adding /eh n d/ as a second dictionary entry for AND, whereas partial pronunciation change requires other solutions. In order to answer these questions we treat the baseform/surface-form pair, *e.g.* (ae,eh), as a unit and analyze the acoustics of such units. The analysis proceeds as follows.

The baseform transcriptions $\hat{\mathbf{B}}$ and surface form transcriptions $\hat{\mathbf{S}}$ (hand labels) of the phonetically labeled training data are first aligned to obtain "pair transcriptions" $\widehat{\mathbf{BS}}$. The following three sets of context independent acoustic phonetic models are then estimated from this set of transcriptions.

- $P_{\mathbf{A}|\mathbf{B}}(\cdot|\cdot)$: estimated from the baseform transcriptions;
- $P_{\mathbf{A}|\mathbf{S}}(\cdot|\cdot)$: estimated from the surface form transcriptions;
- $P_{\mathbf{A}|\mathbf{B},\mathbf{S}}(\cdot|\cdot)$: estimated from the pair transcriptions. [2]

2.1. Acoustics of Alternate Realizations

Our analyses begin by visualizing how the average acoustic features corresponding to an instance of a baseform phoneme /b/ that is realized as a surface form phone [s] compare to those of the baseform /b/ (no matter what the surface form phone is) and the surface-form [s] (no matter what the baseform phoneme is). Note that since we estimate single Gaussian output densities for our acoustic models $P_{\mathbf{A}|\mathbf{B},\mathbf{S}}(\cdot|(\mathtt{b},\mathtt{s}))$, the model mean $\mu_{BS}((\mathtt{b},\mathtt{s}))$ may also be interpreted as a typical acoustic frame when a /b/ is realized as an [s]. We therefore focus attention on the relative location of $\mu_{BS}((\mathtt{b},\mathtt{s}))$ with respect to $\mu_B(/\mathtt{b}/)$, the model for a canonical /b/, and $\mu_S([\mathtt{s}])$, the model for a realization [s].

These means are 39-dimensional vectors (the output of the MF-PLP front-end, and Δ, $\Delta\Delta$ coefficients) which makes visualization difficult. However, since three points in a Euclidean space form a plane, we can find the plane containing the three means and plot them in two-dimensions.

In order to extend this visualization from a single triple for a particular choice of (b,s), say (ae,eh), to a set of triples, we map two points of each triple to two fixed points in the plane and scale the coordinates for the third point such that *relative* distances between the three means are preserved. In particular, $\mu_B(/\mathtt{b}/)$ is mapped to the origin, $\mu_S([\mathtt{s}])$ is mapped to (1,0) on the x–axis, and $\mu_{BS}((\mathtt{b},\mathtt{s}))$ to the positive y–half-plane while preserving relative distances. By plotting model means for all triples in this manner we obtain a plot that gives us the relative location of the three sets of points for different (b,s) pairs. The plot in the center in Figure 1 is obtained in this manner.

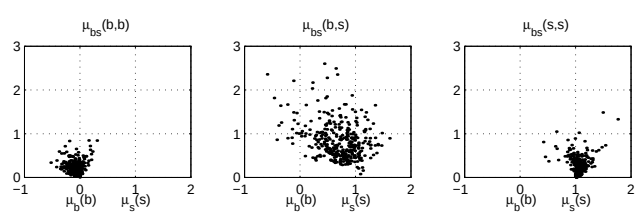

Figure 1: Comparison of average acoustics

To help interpret the results of the plot generated as described

[1] We use the notation /·/ to denote baseform phonemes and [·] to denote surface form phones.

[2] Note that $P_{\mathbf{A}|\mathbf{B}}(\cdot|\cdot)$ and $P_{\mathbf{A}|\mathbf{S}}(\cdot|\cdot)$ have ∼50 HMMs whereas $P_{\mathbf{A}|\mathbf{B},\mathbf{S}}(\cdot|\cdot)$ has ∼500 HMMs (out of ∼2500 possible HMMs – not all possible are seen in our corpus).

above, we also substitute $\mu_{BS}((\mathtt{b},\mathtt{s}))$ with $\mu_{BS}((\mathtt{b},\mathtt{b}))$ for each pair `(b,s)` and obtain the plot to the left in Figure 1. This corresponds to the location of the average acoustic of a `/b/` realized as a `[b]`. Similarly we obtain the location of the average acoustics $\mu_{BS}((\mathtt{s},\mathtt{s}))$ of an `/s/` realized as a `[s]`. These are plotted on the right in Figure 1. 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]` as shown by the plot on the right.

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*: there is a distinct bias in the surface acoustics towards the acoustics of the canonical phoneme. In many instances, the acoustics are actually closer to model for `/b/` than the model for `[s]`!

The conclusions supported by these plots are that

- 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]`;
- neither the phoneme `/b/` nor the phone `[s]` provide a good fit for this realization; and that
- pronunciation change is (spectrally) partial.

2.2. Acoustic Likelihood of Alternate Realizations

In order to see how best to model the acoustics of the `(b,s)` pair, we compare the likelihood assigned to the acoustics by the three models discussed above. For each segment of the acoustics, both in the training set and the 30 minutes of test data, we have the canonical phonemic transcription, the manual phonetic labels, and their alignment (pair labels). The inventories of the canonical and manual transcriptions are identical. In light of this, we compute likelihoods with four model-transcription combinations:

- the canonical pronunciations $\hat{\mathbf{B}}$ with models $P_{\mathbf{A}|\mathbf{B}}$,
- the manual phone labels $\hat{\mathbf{S}}$ with models $P_{\mathbf{A}|\mathbf{B}}$[3],
- the manual phone labels $\hat{\mathbf{S}}$ with models $P_{\mathbf{A}|\mathbf{S}}$ and
- the pair labels $\widehat{\mathbf{BS}}$ with models $P_{\mathbf{A}|\mathbf{B},\mathbf{S}}$.

If we select the instances when a pronunciation change is labeled to have taken place, we find the total likelihoods ordered as

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

Figure 2 summarizes the results of computing these likelihoods for instances in the training data and the test data when a baseform is substituted with another phone.

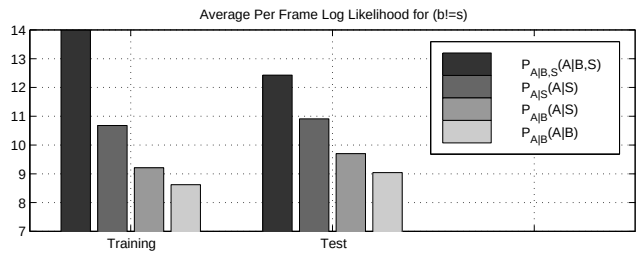

Figure 2: Average per frame log likelihood of training and test data

First of all, we see that $P_{\mathbf{A}|\mathbf{B}}(\cdot|\hat{\mathbf{S}})$, is higher than $P_{\mathbf{A}|\mathbf{B}}(\cdot|\hat{\mathbf{B}})$, although not by much, which helps to explain the moderate gain obtained by pronunciation modeling used only during decoding.

More importantly, $P_{\mathbf{A}|\mathbf{S}}(\cdot|\cdot)$ is better than $P_{\mathbf{A}|\mathbf{B}}(\cdot|\cdot)$. This shows that using surface form transcription is definitely useful for acoustic model training, and demonstrates the value of hand transcriptions.

Finally, the likelihood assigned by the models based on the pairs is significantly better than all the others, despite the over-training indicated by the difference in training and test likelihoods. This clearly shows that it is worthwhile to investigate building acoustic models based on the pairs.

2.3. Temporal Characteristics of Alternate Realizations

So far we have not paid much attention to the fact that the acoustics of a realization of a phone is actually represented by a sequence of feature vectors. In this section we will focus on the sequential or temporal properties of the acoustic realizations. In order to see why the temporal characteristics are important in terms of partial pronunciation change consider the following (for now hypothetical) example where the word `HAD` which has the baseform `/hh ae d/` is labeled as `[hh ae d]` by one human transcriber and as `[hh eh d]` by another. In this case the surface form representation of

[3] This combination provides insight into the case where no change is made to the conventional acoustic model training procedure, but a pronunciation model is used to create dictionary entries for alternate pronunciations.

the baseform /ae/ is ambiguous. Now assume that the first half of the acoustic realization of the baseform /ae/ actually sounds like [ae] but the second half sounds like [eh]. This would explain why the two transcribers disagree. This kind of partial pronunciation change would be better modeled if the surface form representation had higher temporal resolution.

Do such realizations exists in real speech? If they do, how frequent are they? In order to answer these questions we will use the acoustic models estimated from the baseform transcriptions, $P_{\mathbf{A}|\mathbf{B}}$, 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 frames of acoustic vectors. Following this model structure, we investigate the effects of allowing pronunciation changes at each state and at each frame.

In the following experiments, we use context independent acoustic models with single Gaussian state output densities which are trained separately on the baseform and surface-form transcriptions. We focus on instances in the test data when a phoneme is substituted with another phone. For each instance we investigate if allowing pronunciation variation at a subphonetic level would result in higher likelihood. For this purpose we compute the likelihood[4] assigned to the acoustic realization of the phoneme under three models constructed as a combination of the baseform and surface-form models. These models which allow pronunciation changes respectively at entire phone, each state or every frame are illustrated in Figure 3. By construction, all models contain the paths corresponding to Model 1 which corresponds to a total change in the entire phonetic segment.

For each of these models, we then find the best path and count the number of times this best path consists of a sequence of HMM states corresponding to the baseform, surface form or some temporal combination of the two. The percentages computed over all the instances in the test set that correspond to a phone substitution are reported in Table 1.

Model	Best Path		
	$b_1 b_2 b_3$	$s_1 s_2 s_3$	others
Model 1	41%	59%	0%
Model 2	20%	30%	50%
Model 3	15%	23%	62%

Table 1: Distribution of the best state sequence when a canonical phoneme b is realized as a surface phone s.

These results indicate that in a majority of instances of pronunciation change the highest likelihood path corresponds to a temporal combination of the baseform and the surface form. Therefore pronunciation change is (temporally) partial and increasing the temporal resolution of the pronunciation model may be effective in modeling partial changes.

[4]HMM transition probabilities are ignored in this likelihood calculation. Note that for comparison purposes this is equivalent to assuming that all the transition probabilities are equal.

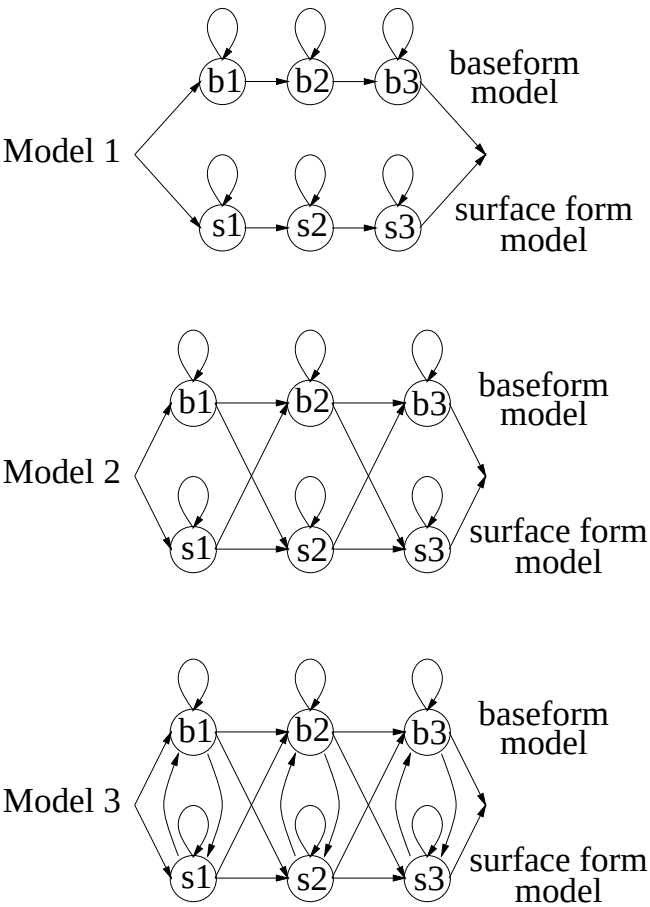

Figure 3: Models used for increasing the temporal resolution

3. Automatic Phone Level Transcription of Acoustic Data

It seems clear from the previous section that training acoustic models with joint knowledge of the canonical and surface form transcriptions may be of significant value. The amount of available hand labeled phone transcriptions for conversational speech is limited, and certainly not enough to build a state-of-the-art state-clustered cross-word triphone ASR system. The phonetic hand labeling process is time consuming and expensive which makes automatic methods for phone transcription desirable.

Our method, described in [7], uses the hand labeled data as bootstrap material. First an initial pronunciation model and

an acoustic model are estimated using the manual transcriptions. Starting with the canonical transcription of the entire acoustic training set, the pronunciation model is used to generate pronunciation networks representing possible phonetic realizations of each training utterance. The most likely phone sequence through each network is chosen via Viterbi alignment using existing acoustic models, yielding a surface form transcription for the entire training set. A new pronunciation model is then estimated using these surface form transcriptions and the above process of network generation and alignment is repeated with this new pronunciation model, giving the final surface form transcriptions.

The quality of the automatic transcriptions is measured by comparing them to the manual transcriptions. This comparison gives a phone error rate (PER) of 26.6% measured on the test set.

At this point it seems natural to ask if the accuracy of this transcription can be further improved upon or if the inherent ambiguity in the acoustics is limiting further improvement in phone recognition. It is also interesting to compare the automatic transcription with the performance of human labelers. A small portion ($\sim$2000 phones) of the hand labeled corpus was transcribed in common by two transcribers and we use this portion to assess interlabeler agreement between human labelers and between our automatic transcription and the human labelers.

Greenberg[5] reports interlabeler agreement on this corpus to be "ca. 75%-80%" [6]. Since the PER using automatic transcription is not so far from the mismatch between the human labelers, it is of interest to examine the performance of the automatic transcriptions with respect to both labelers. This comparison requires a three way alignment and we have done this by hand. An actual example segment of this alignment for the word PARENTS and the overall proportion of each type of agreement is given in Table 2.

			Agreement			Overall
T1	T2	A	T1≡T2	A≡T1	A≡T2	Proportion
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%	72.2%	

Table 2: Example alignment segment and proportion of agreement types

[5] The symbol set used by the transcribers is more detailed than the phone set used in our baseform dictionary (PronLex). Since the rest of our models use the PronLex phone set, we map the actual labels down to this set.

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

If one concentrates on the portion of the transcriptions where the transcribers agree (T1≡T2), the PER is still 14.5% which shows that the disagreement between the automatic and hand transcriptions do not completely overlap with those between the transcribers. It also shows that there is some room for further improvement of the automatic transcription process described here.

The PER between automatic and human transcriptions jumps to >60% in the regions of pronunciation ambiguity, *i.e.* instances where the human transcribers disagree. From a modeling point of view this high error rate is a good reason to keep the surface form representation as a hidden variable during estimation and decoding, in order to alleviate the effects of ambiguity.

4. Speech Recognition Experiments

Two sets of speech recognition experiments have been conducted to evaluate the performance of the acoustic models that are designed to handle partial pronunciation change. Slightly less than 2 hours of speech from the Switchboard corpus make up the test set, of which a 30 minute portion is also phonetically labeled. The baseline acoustic models are state-clustered cross-word triphones trained on canonical phonetic transcriptions of about 60 hours of speech. We use the new acoustic models for rescoring lattices generated by the baseline models. Without *any* pronunciation modeling, the best path in the lattice has a WER of 39.4%.

In the first set of experiments, we use, for a pronunciation model, an explicit listing of the canonical and alternate pronunciations of words in the recognition dictionary (see [1]). We then compare the three models: $P_{\mathbf{A}|\mathbf{B}}$ and $P_{\mathbf{A}|\mathbf{S}}$, which differ in the transcriptions on which they were trained but use the same phonetic inventory, and $P_{\mathbf{A}|\mathbf{B},\mathbf{S}}$, which is trained on the pair transcriptions. The test set, in this case, is only the phonetically annotated portion of the larger test set. The word error rate (WER) measured against the word level transcriptions and phone error rate (PER) measured against both the baseform transcription $\hat{\mathbf{B}}$ and surface form transcription $\hat{\mathbf{S}}$ (hand labeled) are presented in Table 3.

In the second set of experiments, we use a recently introduced method for pronunciation modeling called *state level pronunciation model* or SLPM [7], which accommodates alternate surface-form realizations of a phoneme by allowing the HMM state of the model of the baseform phoneme to share output densities with models of the alternate surface

Acoustic Model	PER wrt $\hat{\mathbf{B}}$	PER wrt $\hat{\mathbf{S}}$	WER
$P_{\mathbf{A}\|\mathbf{B}}$	34.69%	48.10%	48.96%
$P_{\mathbf{A}\|\mathbf{S}}$	42.86%	43.57%	50.57%
$P_{\mathbf{A}\|\mathbf{B},\mathbf{S}}$	33.79%	43.93%	47.81%

Table 3: Recognition performance of acoustic models with a rich pronunciation dictionary

form realizations. The SLPM can effectively "merge" two sets of acoustic models, as described in detail in [7].

We contrast merging the baseline models $P_{\mathbf{A}|\mathbf{B}}$ with the surface-form trained models $P_{\mathbf{A}|\mathbf{S}}$ as described in [7], with the alternative of merging with $P_{\mathbf{A}|\mathbf{B},\mathbf{S}}$, which were shown to better model the acoustics in Section 2. Table 4 shows the results of these experiments.

Note that pronunciation modeling techniques described in [1], which account for complete pronunciation change but not partial pronunciation change, improve the overall WER from 39.4% to 38.9%. Accounting further for the partial change leads to more significant improvements, as seen on the last lines of both tables.

Acoustic Model	PER wrt $\hat{\mathbf{B}}$	PER wrt $\hat{\mathbf{S}}$	WER (Subset)	WER (Full)
$P_{\mathbf{A}\|\mathbf{B}}$	34.58%	50.08%	48.96%	38.8%
$P_{\mathbf{A}\|\mathbf{B}} \cup P_{\mathbf{A}\|\mathbf{S}}$	33.13%	48.48%	47.85%	38.2%
$P_{\mathbf{A}\|\mathbf{B}} \cup P_{\mathbf{A}\|\mathbf{B},\mathbf{S}}$	32.84%	49.37%	47.22%	37.7%

Table 4: Performance of the SLPM on the phonetically annotated subset and the entire test set

5. Concluding Remarks

We have presented acoustic evidence which demonstrates the prevalence of partial pronunciation change in spontaneous conversational speech. We have shown how these partial changes make the notion of phonetic transcription, be it manual or automatic, a difficult one. We have presented means for automatically generating reasonably accurate phonetic transcriptions and a method for using them to train models which improve speech recognition accuracy by accommodating pronunciation ambiguity. A 1.7% WER improvement on Switchboard is demonstrated.

References

1. W. Byrne, *et al*, "Pronunciation Modelling Using a Hand-labelled Corpus for Conversational Speech Recognition", in *Proc. ICASSP '98* Seattle, WA, May 1998.
2. L. Deng, "A Dynamic, Feature Based Approach to the Interface Between Phonology and Phonetics for Speech Modeling and Recognition", Speech Communication (24), pp. 299-323, 1998.
3. L. Deng and D. Sun, "A Statistical Approach to Automatic Speech Recognition Using the Atomic Speech units Constructed from Overlapping Articulatory Features" J. Acoust. Soc. Am. 95(5), pp 2702-2719, May 1994.
4. M. Finke, J. Fritsch, D. Koll and A. Waibel, "Modeling and Efficient Decoding of Large Vocabulary Conversational Speech" in *Proc. EUROSPEECH '99* Budapest, Hungary, Sept 1999.
5. S. Greenberg, "The Switchboard Transcription Project", *1996 LVCSR Summer Workshop Technical Reports*, 1996, `http://www.icsi.berkeley.edu/real/stp/`
6. S. Greenberg, "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, 1998.
7. M. Saraclar, H. Nock, S. Khudanpur, "Pronunciation Modeling by Sharing Gaussian Densities Across Phonetic Models", in *Proc. EUROSPEECH '99* Budapest, Hungary, Sept 1999.