ELSEVIER

Speech Communication 29 (1999) 137–158

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Effects of speaking rate and word frequency on pronunciations in convertional speech ☆

Eric Fosler-Lussier [a,b,*], Nelson Morgan [a,b]

[a] *International Computer Science Institute, 1947 Center Street, Suite 600, Berkeley, CA 94704, USA*
[b] *University of California Berkeley, Berkeley, CA 94720, USA*

Abstract

Automatic speech recognition (ASR) systems typically have a static dictionary of word pronunciations for matching acoustic models to words. In this work, we argue that, in fact, pronunciations in spontaneous speech are dynamic and that ASR systems should change models in accordance with contextual factors. Two variables, speaking rate and word frequency, should be particularly promising for determining dynamic pronunciations, according to the linguistic literature. We analyze the relationship between these factors and realized pronunciations through a statistical exploration of the effects of these factors at the word, syllable, and phone levels in the Switchboard corpus. Both increased speaking rate and word likelihood can induce a significant shift in probabilities of the pronunciations of frequent words. However, the interplay between all of these variables in the realization of pronunciations is complex. We also confirm the intuition that variations in these factors correlate with changes in ASR system performance for both the Switchboard and Broadcast News corpora. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: ASR pronunciation models; Speaking rate; Word predictability

1. Introduction

In recent years, research sites have built experimental Automatic Speech Recognition (ASR) systems for transcription of spontaneous conversational speech. While systems have been successfully marketed recently for dictation with at least 20,000-word vocabularies or recognition of digits or keywords over the telephone, performance on transcription of human-to-human conversations or radio interviews lags behind these systems. Many factors (e.g., channel noise) contribute to the degradation on these tasks; in this study, we present evidence that inadequate pronunciation models are partially to blame. The goal of pronunciation models in an ASR system is to facilitate matching between words and acoustic models.

Because the variability in pronunciations in conversational speech is wider than in other types of speech (e.g., prompted speech), more elaborate pronunciation models are needed to capture the variation.

ASR researchers have found that segmental context can often be an influential factor in capturing some of this variability (Young et al., 1994; Riley, 1991). However, there are other external variables suggested by linguistic studies that can be brought to bear on pronunciation models. In this

☆ Speech files available. See www.elsevier.nl/locate/specom

* Corresponding author. Tel.: +1-510-642-4274; fax: +1-510-643-7684.

E-mail address: fosler@icsi.berkeley.edu (E. Fosler-Lussier)

0167-6393/99/$ - see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0167-6393(99)00035-7

work, we examine two classes of factors: estimates of speaking rate and measures of word predictability. For instance, linguists have recognized that word frequency affects the perception and production of phones (Ganong, 1980; Bybee, 1996 inter alia); speech researchers (Withgott and Chen, 1993 inter alia) have used the concept of *function words* as an approximation to this factor.

Unusually slow or fast speaking rate has also been shown to have an adverse effect on recognizers (Pallett et al., 1994; Siegler and Stern, 1995; Morgan et al., 1997), correlating with increased errors at speaking rate extrema. Linguists have also found that variations in rate of speech can affect both perception and production of phones (Miller and Grosjean, 1981; Summerfield, 1981 inter alia). While each of these factors has individually received attention in the linguistics community, we feel that a study of their joint effect on pronunciations, particularly within the ASR framework, is needed.

It has been noted extensively that fast speaking rates tend to coincide with significant phonological reduction, particularly in spontaneous speech (Bernstein et al., 1992 inter alia). One question we pursue here is whether the phonetic reduction due to fast speaking rate is concentrated in a particular sub-class of words, or occurs across-the-board in the vocabulary. From an information-theoretic perspective, speakers would preserve the most information by concentrating reductions in the least informative words – that is, the words most predictable from context. Thus, we hypothesize that there is a significant interaction between rate and word predictability.

The implication of this study for ASR pronunciation models is that if speaking rate and word predictability are interacting features that can help predict when pronunciation variability is expected, then pronunciation models within the ASR system should change dynamically in response to these variables. We would expect changes in speaking rate, as well as the predictability of words, to correlate with both deviations from the "normal" ASR pronunciation models and the accuracy of the recognizers that employ these models.

This hypothesis assumes that the pronunciation model plays a significant role in the recognition of spontaneous speech – an assumption supported by the recent work of McAllaster et al. (1998). Using a simulated data approach to evaluate their recognizer, they found that the pronunciation dictionary was a significant source of errors in their spontaneous speech system:

> Put most provocatively, the variant and reduced pronunciation of casual speech accounts for most of the errors made by this recognition system. (McAllaster et al. 1998, p. 1848)

In this study, we give additional evidence to support this claim by studying the effects of deviation of pronunciations from ASR pronunciation models on the performance of the system. In our first recognizer analysis, we used a phonetically hand transcribed corpus to determine actual word pronunciations. However, as many automatic pronunciation learning systems use recognizer acoustic models to generate pronunciation alternatives, we also analyzed phonetic transcriptions based on a phone recognizer.

In the ASR context, it is important to pinpoint when deviations from dictionary pronunciations can occur. Phonetic reduction processes tend to make the pronunciations of different words more similar, which can be devastating to ASR systems in terms of word confusability if all possible pronunciations of every word are added indiscriminately to the lexicon. Thus, an investigation of how these factors affect pronunciations will benefit ASR models that judiciously allow variations to occur in the dictionary.

In summary, there are several questions our work tries to answer:

- Are there systematic trends in pronunciation variation with respect to rate and frequency?
- What is the effect on ASR systems when the actual pronunciation is different from what the recognizer expects?
- Do the factors we have chosen (speaking rate and word predictability) have an effect on recognizer performance for spontaneous speech systems? If so, what is the interaction with pronunciations?

- Do the results for phonetic hand-transcription analysis carry over to automatically generated phone transcriptions?

In Section 2, we briefly discuss ASR pronunciation models and how our proposal for the inclusion of speaking rate and word predictability can be thought of as an extension of the concept of context, already present in many automatic learning systems. Section 3 discusses how we estimate speaking rate and word predictability. Section 4 presents an analysis of pronunciation variation in phonetic hand transcriptions of the Switchboard corpus at the word, syllable, and phone levels. We analyze an ASR system trained on the Switchboard corpus and the effects of pronunciation variation upon its performance in Section 5. In Section 6, we examine another recognizer trained on the Broadcast News (BN) corpus, in order to compare the effects of the pronunciation model across corpora. Finally, we present our conclusions in Section 7.

2. Related work

As we stated above, the goal of this research is to examine how word predictability and speaking rate influence word pronunciations, with the eventual goal of using these factors to influence dynamic ASR models of pronunciation. We view these variables as facets of the contextual environment in which the word exists. Obviously, other types of context can be used in pronunciation models; in this section, we summarize several different techniques.

Most rule-based pronunciation models, using either decision trees (e.g., Chen, 1990; Riley, 1991) or phonological rules (e.g., Tajchman et al., 1995; Finke and Waibel, 1997b), do use context to constrain pronunciations to some degree. In these paradigms a baseform dictionary, which gives standard "canonical" pronunciations, is expanded into a set of alternative pronunciations (or *realizations*) for each word. The transformation procedure typically uses a number of surrounding baseform phones to determine possible realizations for a particular phone. Such techniques have improved ASR systems for Switchboard over static lexicons by about 1–2% (Finke and Waibel, 1997b; Weintraub et al., 1997; Riley et al., 1998).

Other systems have used the surrounding words as the contextual influence in determining baseform pronunciations. For instance, Sloboda and Waibel (1996) built pronunciations for pairs and triples of German words in the Verbmobil task. Gauvain et al. (1997) and Placeway et al. (1997) included pronunciations for frequent pairs of words in their BN recognizers. None of these papers relates how much improvement is attributable to multi-word modeling. However, in the Switchboard domain, Riley et al. (1998) showed that incorporating multiwords into the dictionary and reestimating pronunciation probabilities reduced error by about 1% absolute (3% relative) over the Pronlex dictionary (LDC, 1996) on the Switchboard corpus.

However, the neighboring phones and words are not the only contextual factors that can be used for determining pronunciation variants. Ostendorf et al. (1997) and Finke and Waibel (1997b) have used several extra-segmental features, such as rate of speech and pitch, to determine a hidden speaking mode (a statistical clustering of pronunciation variations) in order to influence the pronunciation models of individual phones. We (Fosler-Lussier and Williams, 1999) have also constructed models that incorporate segmental factors, speaking rate, word predictability, and other factors (e.g., time since the previous pause) to determine syllable and word models, showing a 1.4% absolute (5% relative) gain over static models on the spontaneous portion of the BN corpus.

2.1. Why is context important?

The main function of context within an ASR pronunciation learning system is to limit the number of possible pronunciations in order to curb word confusability within the recognizer. The need to constrain the number of pronunciation alternatives was shown in a diagnostic study by Saraclar (1997). In this study, Switchboard lattices (baseline: 46% word error) were rescored in two ways. In both experiments, a phone constraint decoding on the test set was used to determine the

word pronunciations that matched the speech recognizer's acoustic models. These new word pronunciations were used to supplement the recognizer dictionary. In the first trial, for each utterance, the dictionary was augmented with the new pronunciations corresponding to just that utterance. Using the resulting lexicon to rescore the lattices reduced the word error to 26%. However, when the lattices were rescored using a dictionary containing the new pronunciations for *all* of the test set, the word error rate increased to 38%. In other words, the benefit of having the correct pronunciation was often offset by the presence of unnecessary competing pronunciations. This result illustrates the importance of dynamically selecting appropriate pronunciations.

3. Measures of speaking rate and word predictability

In order to determine how extra-segmental factors can affect word pronunciations, one must first determine a set of measurements for these factors. In this section we discuss several measures of the rate of speech and the predictability of words that were used in this study.

3.1. Speaking rate

Speaking rate is generally measured as a number of linguistic units per second, although the actual units used are open to some question. Previously, we have shown that using units other than words per second (e.g., phones per second) as a metric is more reliable for predicting word error in ASR systems (Mirghafori et al., 1995). Contrarily, Fisher (1996a) showed that, for the Hub3 North American Business News task, the difference between words and syllables per second was insignificant for prediction of recognizer error rate. However, Fisher preferred the syllabic measure because it was likely to be easier to calculate independent of any recognizer.

For this study we chose the syllabic rate as our metric. Syllables are far less likely to be deleted than phones; in the Switchboard corpus, the phone deletion rate is roughly 13% (Weintraub et al., 1997), whereas complete phonetic deletion of the syllable occurs only 2.5% of the time. Since we are trying to predict phone deletions in our pronunciation models, syllabic rate is a more stable measure for our purposes.

There are several ways that syllabic rate can be determined from speech data. In this study we use *transcribed syllable rate*, which is determined from syllabic boundaries notated by linguistic transcribers. This interpausal rate is determined by counting the number of syllables between transcribed silences and dividing by the amount of time between pauses. This particular measure is generally not determinable at recognition time for interactive systems. Nonetheless, as speaking rate estimators can sometimes be unreliable (particularly for spontaneous speech), we used this metric as an incontrovertible measure for determining the effects of speaking rate on pronunciations.

When syllabic annotations are not available, one can also determine asyllabic rate from the alignment of the word transcription to speech data (*aligned syllable rate*); since the syllable deletion rate is roughly 2.5% and the insertion rate is negligible, this corresponds closely to transcribed rate.

However, at the run-time of the recognizer other metrics must be used, as the above measures are not feasible to calculate. In (Mirghafori et al., 1996), we described the tactic of running the recognizer twice, using the first pass to hypothesize sound unit boundaries and hence the speaking rate, which would then be incorporated in a second pass. Yet, aside from the additional computation, this method requires the assumption that the speaking rate determined by a potentially erroneous recognition hypothesis is sufficiently accurate. For difficult tasks such as conversational speech recognition this is often not the case, particularly for unusually fast or slow speech.

Alternatively, one can use signal processing or classification techniques to estimate speaking rate directly from the acoustic signal (Kitazawa et al., 1997; Verhasselt and Martens, 1996 inter alia). We have also derived such a measure, dubbed *mrate* for its *m*ultiple *rate* estimator components (Morgan and Fosler-Lussier, 1998). The measure correlates moderately well with transcribed syllable

rate ($\rho \sim 0.75$), although it tends to underestimate the rate for fast speech.

3.2. Word predictability

The most obvious candidate for determining word predictability is the unconditional probability of the word (i.e., P(word)), determined from the number of instances of the word in the reference transcription of the entire corpus. This is known in ASR parlance as the *unigram probability* of the word. The results reported here use the base 10 logarithm of the unigram probability.

However, the predictability of a word in context may also have an effect on its pronunciation. One simple measure of the localized predictability used by ASR systems is the *trigram probability* ($P(\text{word}_n|\text{word}_{n-2}, \text{word}_{n-1})$) – the probability of the word given the previous two words. In the case where the trigram was not available, a Good-Turing backoff strategy was employed, in which the trigram is estimated from a bigram ($P(\text{word}_n|\text{word}_{n-1})$) and a weighting factor, as is done by several ASR systems. While we could have chosen to examine bigram scores (one of the ASR systems we evaluate later uses a bigram, the other a trigram), we wanted to include as much contextual information as possible in order to distinguish this measure from the unigram score.

One can also imagine more elaborate models, such as semantic triggers, word collocations, or syntactic constraints, that could be used to predict when a word is more likely. However, for the sake of this study, we chose to only use models conveniently available to most speech recognizers.

4. Relationships between speaking rate, word predictability and pronunciation changes

In this section we present statistical analyses that show the relationship between dynamic factors and pronunciations in the Switchboard corpus (NIST, 1992), a collection of telephone conversations between two strangers in which speakers were asked to talk about one of hundreds of topics and were recorded for up to five minutes. We begin with an analysis of how pronunciations of individual words deviate from the canonical.

4.1. Experimental design

The speech data from the Switchboard corpus used for this study are a subset of the complete database, consisting of approximately four hours of phonetically hand transcribed utterances provided by ICSI for the Johns Hopkins Summer Research Workshop series (Greenberg, 1997). About one half hour of this data was from the development test set, while the rest was from the training set. Starting from an automatic syllabic alignment generated by the Johns Hopkins' HTK recognizer, linguists from ICSI realigned the syllable boundaries and identified syllables with their phonetic constituents. [1]

Using an automatically syllabified version of the Pronlex dictionary [2] (LDC, 1996), we generated a mapping from dictionary baseforms to these hand transcriptions using a dynamic programming technique developed by Weintraub et al. (1997). The procedure uses a string-edit-distance algorithm, where the distance metric between two phones ϕ and ψ is given by

$$d(\phi,\psi) = \sum_{f \in \text{Features}} g(f(\phi), f(\psi)), \tag{1}$$

$$g(f(\phi), f(\psi)) = \begin{cases} 0 & \text{if } f(\phi) = f(\psi), \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

The features of each phone are a set of 24 binary phonetic features (e.g., front, nasal, high) similar to those found in (Chomsky and Halle, 1968). Thus, the phones [f] and [v], which differed in two features,

[1] See also http//www.icsi.berkeley.edu/real/stp for more information about the Switchboard Transcription Project.

[2] The Pronlex dictionary was syllabified at the 1996 Johns Hopkins Summer Research Workshop (WS96) using Fisher's (1996b) automatic syllabification program, which is based on Kahn's (1980) thesis. This dictionary is used by several ASR systems (without syllabification), including the recognizer we analyze in Section 5. Thanks to Barbara Wheatley and others at WS96 for help with this lexicon.

$$[\mathtt{f}]:\begin{bmatrix}+\text{tense}\\ -\text{voiced}\end{bmatrix},\quad [\mathtt{v}]:\begin{bmatrix}-\text{tense}\\ +\text{voiced}\end{bmatrix},$$

had a distance score $d([\mathtt{f}],[\mathtt{v}])$ of 2, whereas the distance score between [f] and the more dissimilar vowel [ae] was much higher (11).

Every baseform (dictionary) phone was mapped to zero or more hand-transcribed phones; deletions caused the baseform phone to be mapped to zero phones, and insertions caused the dictionary phone to be mapped to multiple transcription phones. In the cases where multiple pronunciations existed in the dictionary [3], the closest baseform (in terms of the distance metric d) to the realization was used.

The output of the alignment procedure was a map α; each instance of a baseform phone ϕ in the database was mapped to an n-tuple of realized phones:

$$\alpha(\phi) = \langle \psi_1, \psi_2, \ldots, \psi_n \rangle. \tag{3}$$

Typically, n was between 0 and 2.

We created these pronunciation maps for every baseform phone in the transcribed database. We then annotated every word (and its syllabic and phonetic constituents) with our measures of speaking rate and word predictability, namely transcribed syllable rate of the region, unigram frequency of the word, and trigram probability in the utterance context. Thus, we obtained a database of pronunciation variation for every word, syllable, and phone in the transcribed portion of the Switchboard corpus.

4.2. Pronunciation difference metrics

One of the difficulties we encountered when embarking upon this study was characterizing the behavior of pronunciations as a function of the factors we would like to study. We experimented with a number of metrics; each has some advantages and disadvantages.

4.2.1. Probability of a single pronunciation

We tracked the probability of canonical pronunciations, which is particularly useful when estimating how well the pronunciations given to the baseline recognizer match the transcribed data. Canonical is defined here as matching a listed pronunciation from the ASR dictionary. For spontaneous speech the canonical and most frequent pronunciations often differ, so we tracked the behavior of the single most frequent pronunciation as well, assuming that a system that performs automatic baseform learning would also have that particular pronunciation in its dictionary.

Using this metric, we are able to tell when the probability of a particular pronunciation has changed significantly [4] due to a change in one of our features. However, a drawback to this metric is that analysis becomes more difficult when tracking more than just a few pronunciations.

4.2.2. Entropy

This is a traditional measure for pronunciation learning systems (see, e.g., Riley, 1991), and is a good measure of the spread of pronunciations in a training set. For a set of pronunciations X with a probability distribution estimate p, the entropy $H(X)$ in bits is defined as

$$H(X) = -\sum_{x \in X} p(x) \log_2(p(x)). \tag{4}$$

This measure becomes unwieldy, however, if one tries to use it to predict the relative entropy of a particular test set – pronunciation models are typically pruned to some cutoff (assigning zero probability to some test events), which causes relative entropy to approach infinity. Simple measures of entropy also treat all pronunciations as distinct and unrelated. Thus, a pronunciation distribution for *this* $\langle p([\mathtt{dh\ ih\ s}]) = 0.8, p([\mathtt{d\ ih\ s}]) = 0.2\rangle$ and a second distribution $\langle p([\mathtt{dh\ ih\ s}]) = 0.8, p([\mathtt{dh}]) = 0.2\rangle$ have the same

[3] The Pronlex dictionary has mostly single pronunciations per word; the average number of pronunciations per word for the 22K lexicon was 1.07, whereas for the 100 most frequent words in Switchboard, this average is 1.14.

[4] When significance is reported here, we mean that two distributions are significantly different at $p \leqslant 0.05$ using a difference of proportions test.

entropy, although the second distribution intuitively seems farther from canonical.

4.2.3. Phonetic distance score

We also developed a metric that was smoother than the hard binary decision of whether a pronunciation was canonical or not by using the phonetic feature distance (d) between the two pronunciations as utilized in the dynamic programming technique described in Section 4.1. For example, the formula for the distance score between two syllables, σ_{base} and $\sigma_{\text{transcribed}}$, depended on each phone ϕ_{base} of the syllable σ_{base}, and the phonetic alignment α from Eq. (3),

$$D(\sigma_{\text{base}}, \sigma_{\text{transcribed}}) = \sum_{\phi_{\text{base}} \in \sigma_{\text{base}}} d(\phi_{\text{base}}, \alpha(\phi_{\text{base}})), \quad (5)$$

where $\alpha(\phi_{\text{base}})$ returns the aligned transcription phones. [5]

We interpret this distance as a measure of how far the realized pronunciation has deviated from the expected pronunciation. Rather than discretizing pronunciations as is done in the entropy and single probability measures, this score integrates the distance between phonetic features associated with each string of phones. This procedure can also be extended to give a smoothed score using a particular pronunciation model; the distance between each baseform pronunciation in the model and the target phone sequence is weighted by the probability of the baseform pronunciation. However, as this measure is not a probabilistic quantity, it is difficult to give a statistical or information-theoretic interpretation to this metric.

4.3. Canonical pronunciations of individual words

In a pilot experiment to show the effects of our features on a coarse level, we extracted the word-pronunciation pairs for the 117 most frequent words from a two hour subset of the transcriptions from the training set. Each word had at least 40 occurrences in the set. For every selected word, we divided the pronunciation population into two halves: words above the median speaking rate and words below the median speaking rate, giving two pronunciation distributions. We compared the probability of both the most likely transcribed pronunciation and canonical pronunciation (as given in the Pronlex dictionary) between partitions. A sample comparison for the word "been" is shown in Table 1.

In this case, the probability of the canonical pronunciation [b ih n] drops significantly for the faster half of the examples. The distribution of alternate pronunciations changes as well: the reduced-vowel variant, [b ix n], only occurs in the fast speech examples. The significant change in probability for the canonical pronunciation was a common occurrence in the top 117 words; we found that there was a significant ($p < 0.05$) shift in canonical pronunciation probability for 30% of the words due to rate differences. For speaking rate differences that were significant, faster rate indicated fewer canonical pronunciations, without exception.

We repeated the partitioning using trigram scores of the words as the splitting criterion. In Table 2, we show the number of words with significant changes in pronunciation probability due to each factor. When analyzed in terms of the trigram, 18% of the words had a significant shift in canonical pronunciation probability. Similar results were seen with the most likely pronunciations.

Table 1
Distribution of the pronunciation probabilities for 45 realizations of the word "been"

Pronunciation	Low syllable rate	High syllable rate
Canonical	0.6087 b ih n	0.3636 b ih n
Alternatives	0.1304 b eh n	0.1818 b ix n
	0.0870 b ih nx	0.1364 b ih nx
	0.0435 b ih n n	0.0909 b ih
	0.0435 b eh n	0.0909 b eh n
	0.0435 b eh nx	0.0455 b eh
	0.0435 b ih	0.0455 b ah n
		0.0455 v ih n

[5] Technically, α returns an n-tuple of phones, but here we extend the interpretation of the distance metric d to include the concept of insertions and deletions: for each insertion or deletion, the insertion/deletion penalty distance used in the alignment procedure is added to the score total.

Table 2
The number of words where significant pronunciation changes were seen based on syllable rate and language model probability for the most frequent words in Switchboard. Pronunciation changes were calculated for the canonical pronunciation and most likely surface pronunciation for each word

Number of words (out of 117) with significant pronunciation changes		
	$p < 0.05$	$p < 0.01$
Canonical prons, SylRate	35 (29.9%)	12 (10.3%)
Canonical prons, LM	21 (17.9%)	5 (4.3%)
Most likely prons, SylRate	31 (26.5%)	12 (10.3%)
Most likely prons, LM	20 (17.1%)	7 (6.0%)

As with speaking rate, higher trigram probability (i.e., if the word was more likely) also meant a decrease in canonical pronunciation probability. It is noteworthy that the words that showed significant difference were often distinct between the rate and language model lists; for $p < 0.05$, only 9 words were in both, meaning that in total 40% of the words had significant shifts due to either rate or trigram score. We could find no clear-cut rationale for why these two lists were mostly distinct.

Probability shifts in the canonical pronunciation more often were due to speaking rate than to word predictability; this is understandable as the distribution of words we examined is already skewed with respect to trigram scores. In order to get enough data for per-word scores, we chose the most frequent words, which are a priori more likely to have higher trigram scores – the range of scores for these words is smaller than that for the general population of words.

Thus, using a relatively gross measure of pronunciation change, we were able to find interrelations between dynamic changes in word pronunciations and our factors. However, the 117 words we examined only covered 68% of the corpus. In order to better characterize pronunciation variation in a wider cross-section of the corpus, we decided to look at pronunciation statistics for individual phones and syllables, for which we had more data.

4.4. Studies of phone realizations

For each dictionary phone, we extracted the corresponding hand transcribed phones, along with the applicable speaking rate. We then observed the overall trends for all of the phones.

As seen in Fig. 1, we found that from very slow to very fast speech the phone deletion rate rises from 9.3% to 13.6%; the phone substitution rate also changes significantly ($p < 0.05$), rising from 16.9% to 24.2%. We found that as speaking rate increases, the entropy of the distribution of phone pronunciations also increases (Fig. 2). A further examination of the data partially explains the entropy increase: as speaking rate increases, phones are not just switching their form from canonical to one single alternate. Rather, phones are being realized with more forms in fast speech. For the slowest speech, the average phone had 3.4 different corresponding realizations occurring at least 2% of the time, whereas in fast speech, phones had an average of 4.0 realizations. When only counting transcriptions that appeared 10% of the time, fast speech still had more realizations (1.9) than slow speech (1.5).

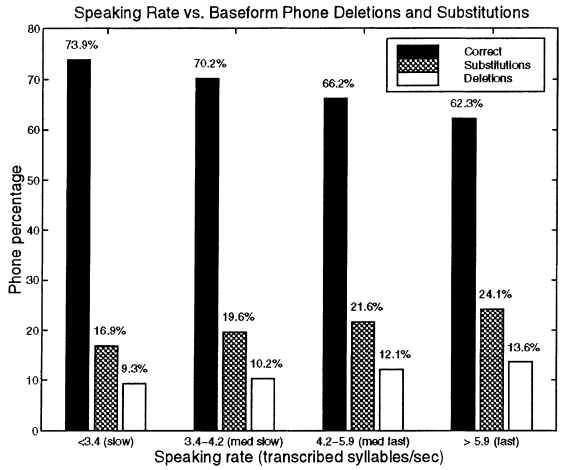

Fig. 1. Phone-level statistics for effects of speaking rate on pronunciations.

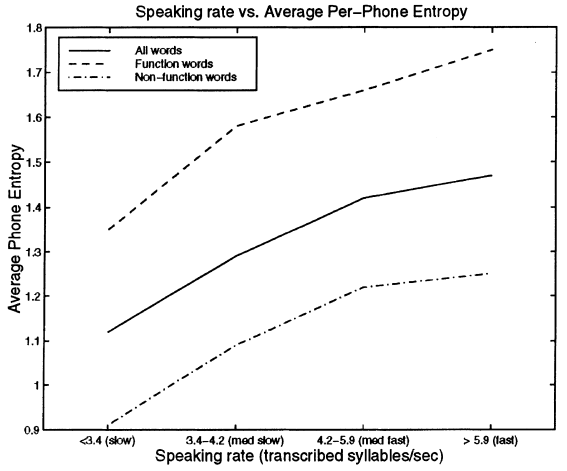

Fig. 2. Phone-level entropy for different speaking rates. The 100 most frequent words were labeled as function words.

From Fig. 2 we can also see that an interaction occurs between word frequency, rate, and phone pronunciation: when the 100 most frequent words are separated out from the general population, we can see that both function and non-function words have similar entropy curves as speaking rate changes. However, function words as a whole demonstrate wider variation in phone pronunciation than words containing semantic content. One might hypothesize that the high frequency of these words contributes to this disparity. However, it is also possible that, since function words are shorter in general, the average pronunciation deviation per phone is higher, but per-word deviation is similar to non-function words. We will revisit this hypothesis in the next section.

4.5. Syllable-level investigations

In our final analysis of the Switchboard hand-transcriptions, we examined pronunciation statistics both for groups of syllables and individual syllable types in the entire four hour transcription set. This allowed us to cluster some of the data from the word-level experiments and permitted us to evaluate the effects on individual phones within their syllabic contexts. It has been suggested that pronunciation phenomena are more often affected by syllabically internal rather than external context (Greenberg, 1998).

4.5.1. Statistics for groups of syllables

We computed the average syllabic distance (Eq. (5) in Section 4.2) for all of the syllables in the set, and plotted them against the unigram frequency of the word and speaking rate. As can be seen from Fig. 3, there is an interaction between unigram probability, speaking rate, and the average distance for each syllable from the Pronlex baseforms: in less frequent words there is some increase in mean distance as rate increases, but for syllables occurring in more frequent words the rate effect is more marked. This complex interdependency between these three variables makes sense from an information-theoretic viewpoint – since high-frequency words are more predictable, more variation is allowed in their production at various speaking rates, as the listener will be able to reconstruct what was said from context and few acoustic cues.

Earlier, we posed a question about the relationship between phone entropy and function words: is the increase in average phone entropy due to the shorter length of function words? Here, we see that the length of function words is not a factor; if it were, then one would expect there to be no difference (or even lower scores) with the

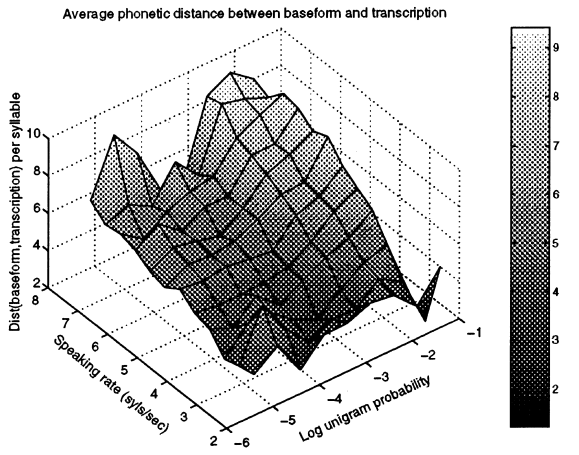

Fig. 3. Average syllable distance from baseform as a function of speaking rate and unigram probability. For low unigram probability and very high or low speaking rates, the number of samples is low; the peaks in the graph for these extremes are probably statistical noise. A color version of this figure is available at www.elsevier.nl/locate/specom

syllabic difference metric, which is unnormalized with respect to the number of phones.

When we replaced the syllable distance metric with the probability of canonical pronunciations metric for this same data we did not observe this general interaction between the metric, unigram frequency, and speaking rate, which puzzled us greatly. The first key to solving the puzzle was to notice that the probability of canonical pronunciations *did* change as a function of rate when we took lexical stress (as marked in Pronlex) and syllabic structure into account. We annotated syllables with `O` for onset consonants, `N` for nuclei, and `C` for codas, repeating symbols for clusters. Thus, `OONC` represents a syllable with a 2-phone onset cluster, a nucleus, and a single-phone coda (e.g., *step*). For each syllable type and stress type (primary, secondary, and none), we calculated the probability that syllables of that type were pronounced canonically, as a function of rate (Fig. 4).

Most of the function words were marked in the dictionary with secondary stress rather than primary stress; therefore, the secondary stress category is somewhat like a function word category in this analysis. This is supported by the fact that the average number of variants per secondary-stressed syllable in this database is 5.3, versus 1.8 alternatives for primary-stressed syllables and 3.2 for unstressed syllables. Looking across columns for each syllable type, these data also confirm that syllabic stress is an important factor in pronunciation models, as has also been observed by other researchers (Finke and Waibel, 1997a; Ostendorf et al., 1997; Weintraub et al., 1997). There also seems to be a trend for syllables without codas to be pronounced canonically more often, as can be seen by (for example) comparing `OONC` and `ONC` to `OON`. This corresponds well with the fact that coda consonants are more frequently changed from canonical (usually by deletion) than onsets in this database, as reported by Greenberg (1998, Table 6).

For some syllable types, (e.g., primary stressed nucleus-only), rate has a strong effect on whether the syllable was pronounced canonically, but for others the effect is negligible. For one case (secondary stressed nucleus-only), [6] a surprising reverse effect occurs – the probability of canonical pronunciation increases as rate increases. Thus, stress and syllabic structure do interact with speaking rate in terms of syllable pronunciations.

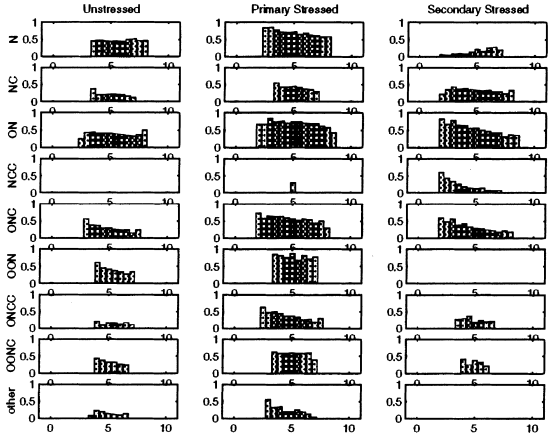

Fig. 4. Probability of canonical pronunciations (y-axis of each subgraph) for different speaking rates (in syllables/s on x-axis), partitioned by lexical stress and syllabic structure. Light grey bars have 20–100 samples, medium grey 100–1000, dark grey >1000. `O` = onset, `N` = nucleus, `C` = coda; multiple occurences of `O` or `C` indicates consonant clusters. Empty boxes indicate no (or little) data. A color version of this graph is available at www.elsevier.nl/locate/specom

4.5.2. Individual syllables

We then examined the 200 most frequent syllables in the Switchboard corpus, which provides 77% syllable coverage of the four-hour transcription set, and 75% of the corpus at large. We clustered the data for each syllable into speaking rate histogram bins, and determined the probability of the canonical and most likely pronunciations [7] for each syllable as a function of the rate bin. We reclustered data in a similar fashion using trigram probability as the clustering criterion.

For every histogram bin, we observed the percentage of syllables that had either the canonical or most-likely pronunciation. Fig. 5 illustrates how

[6] There are only two words in the dictionary that fall into this category: *a* [ey] and *uh* [ah].

[7] The canonical and most likely pronunciations differed for 55 of the 200 syllables; for example, *don't* ([d ow n t]) was most frequently transcribed as [dx ow] (i.e., with a dental flap and deletion of the coda consonants).

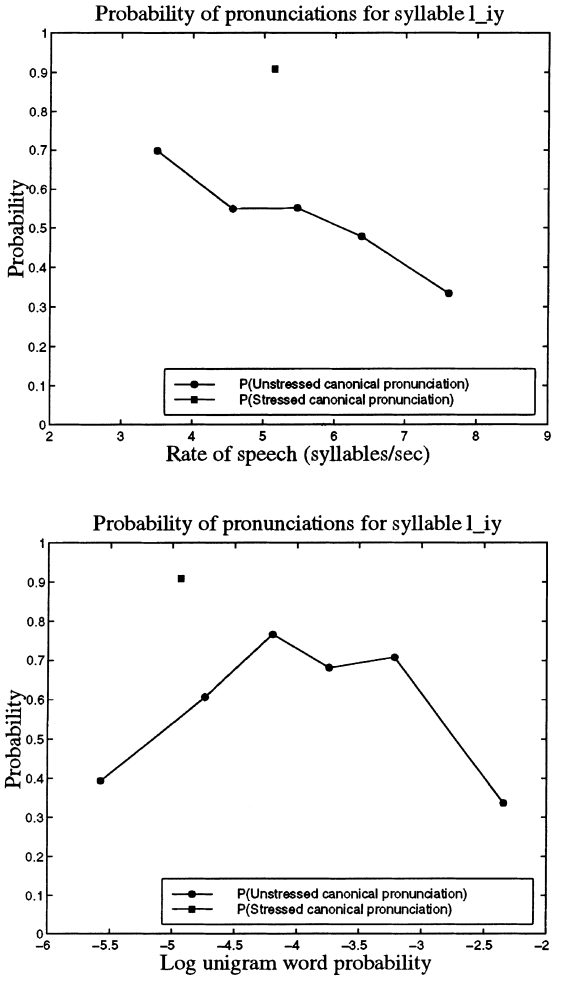

Fig. 5. Pronunciation probabilities for the syllable [l_iy] dependent on rate and unigram frequency.

the pronunciation probabilities change as a function of rate and word frequency for the syllable [l_iy]. There is a significant movement towards alternate pronunciations at faster speaking rates. Unigram frequency of the containing word also has a distinct effect on this syllable – it is least often canonical in the extremes of this metric. We are uncertain why this relationship is non-monotonic in this instance; while this is a rare occurrence, there are several other syllables in our database that exhibit this behavior. On a side note, we also see in these graphs the influence of stress on canonicity: the stressed versions of [l_iy] appear unadulterated much more frequently than the unstressed versions.

For each of our 200 syllables, we then determined whether the probability of the canonical or most-likely pronunciation changed significantly between any two histogram bins as speaking rate or trigram probability varied. As we see in Table 3, changes in the rate of speech significantly affected the probability of the canonical pronunciation in 85 of the syllables; when the most-likely pronunciation is considered as well, the number of affected syllables increases to 95. The trigram probabilities of the word influences fewer canonical pronunciations, although roughly one-third of syllables are still affected.

The major characteristic that describes the class of syllables with significant rate shifts is that these syllables are often more frequent. The mean unigram log (base 10) probability for these syllables is -2.33; for non-affected syllables the mean unigram log probability is -3.03. Thus, the syllables that experience pronunciation changes as a function of rate are generally part of the more frequent words. This is consistent with the earlier syllable distance results in Fig. 3, which showed a more marked effect of speaking rate in syllables appearing in words with a high unigram frequency.

For some syllables (Fig. 6), there is a tradeoff between the most-likely and canonical pronunciations as a function of rate. However, this tradeoff

Table 3
Number of syllables (out of 200) with significant ($p < 0.05$) differences in pronunciation probabilities for the extremes of speaking rate and trigram probability

	# of syls w / significant differences			
Clustering on:	Canonical	Most likely	Either	Both
Speaking rate	85	81	95	71
Trigram probability	64	59	70	53

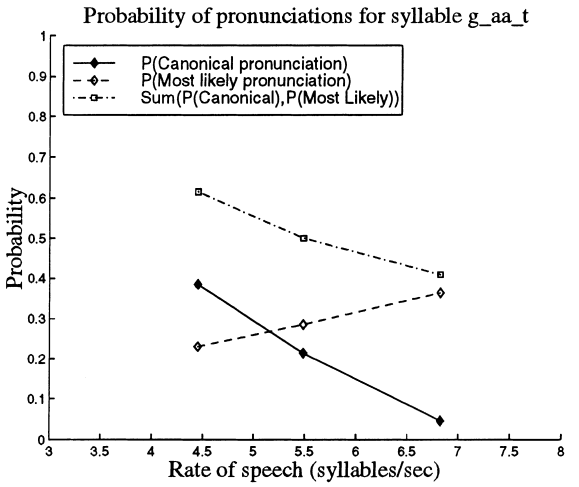

Fig. 6. Canonical [g_aa_t] versus most-likely [g_aa] for varying rates.

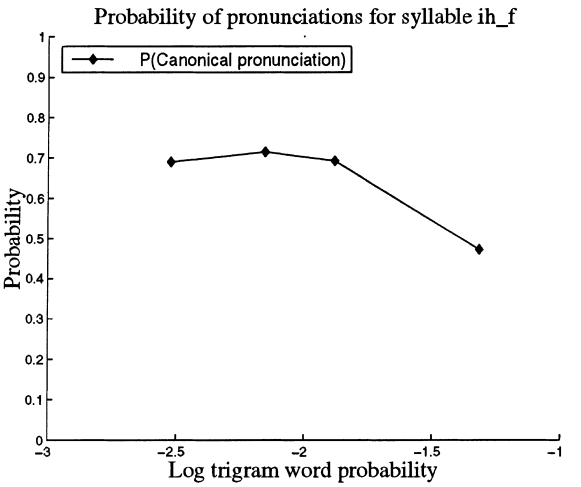

Fig. 7. Probability of canonical [ih_f] pronunciation for various trigram probabilities.

is not completely one-for-one: the sum of the canonical probability and most-likely probability is lower for faster examples than for slower examples. In faster speech, other pronunciations receive more of the probability mass.

Up to this point, we have been treating unigram and trigram scores as roughly equivalent. For the vast majority of cases, it appears that using trigram scores provides little extra modeling power, as the trigram is often correlated with the unigram and the trends in pronunciations often match for the two features. However, for a small number of frequent syllables it distinctly helps to have the trigram score. For example, in Fig. 7, the syllable [ih_f], which corresponds only to the word *if* in our training set (i.e., all examples share the same unigram probability), is significantly reduced in very likely word sequences. In this case, the trigram score supplies extra information for forecasting reductions that the unigram does not provide. Further evidence that the trigram is an effective tool for predicting reductions in high frequency words is presented by Jurafsky et al. (1998); using regression models, they found that trigram probabilities were a significant factor in prediction of word length for six of the ten most frequent words. Trigrams also were useful for predicting change in vowel quality (i.e., whether the vowel was canonical, another full vowel, or reduced) for six of the ten. [8] Thus, this component, already present as the language model in many ASR systems, may be useful for predicting pronunciation change in frequent words.

4.6. Summary

We have analyzed phonetic transcriptions of the Switchboard corpus in order to ascertain the effect of speaking rate and two measures of word predictability on pronunciations of words, syllables, and phones. One of the most significant findings is that not every linguistic unit is affected by changes in these factors. An increase in transcribed syllable rate is correlated with deviation from dictionary baseforms in roughly half of the syllables and a little less than a third of the words we studied. High word predictability (both using the unigram and trigram metrics) also tends to accompany lower canonical pronunciation probabilities, although we have also observed for some syllables that lower canonical pronunciation probabilities can be found in infrequent situations.

We have also seen that there is a significant interaction between the investigated features and

[8] Only one of the 10 words was unaffected in either the length or vowel quality categories.

pronunciations. In particular, Fig. 3 shows that word frequency has a distinct influence on how much pronunciation variation is present with changes in transcribed syllable rate: syllables in high frequency words are most affected by the rate of speech. Stress and syllable structure also play an important part in cooperation with these features; Fig. 4 illustrates that variations due to rate are more visible when these factors are included.

5. Switchboard recognizer error analyses

It is clear from our previous analysis that speaking rate and word predictability are both distinctly correlated with pronunciation change. In this next section, we investigate the effects of pronunciation model mismatches with the Switchboard hand transcriptions on ASR word error rates, and how these mismatches correlate with speaking rate and word predictability.

For these analyses of recognition performance on the Switchboard corpus, we used the HTK recognizer trained with the Pronlex dictionary developed at Johns Hopkins 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop (hereafter referred to as the WS96 recognizer) to provide recognition hypotheses for error analysis. This Hidden Markov Model (HMM) recognizer is a 12-mixture state-clustered cross-word triphone system, trained on 60 hours of mel cepstrum features (including first and second derivatives of the features). The recognizer uses a bigram language model trained on 2.1 million words of Switchboard transcripts.

5.1. Mismatch in pronunciations as a correlate of error

Previous studies (Weintraub et al., 1997) have shown that when the Switchboard hand-transcriptions were compared to the Pronlex dictionary, only two-thirds of the dictionary phones matched the transcriptions. In an elaboration of this study, we have tried to characterize the effects of these phone-level statistics on word-level pronunciations. We found that while 67% of phones retained "canonical" form in spontaneous speech, only 33% of word pronunciations found in the Switchboard development test set (using ICSI hand transcriptions) were found in the Pronlex dictionary. [9] Thus, the phone transformations we have observed are not concentrated in a few words, but rather are spread throughout the corpus.

What remains to be shown is that these pronunciation errors have an effect on our recognizers. Intuitively, one would believe that recognizers would fail miserably if 67% of hand-transcribed word pronunciations are not in the dictionary. However, it is not necessarily true that ASR acoustic models are modeling the same linguistic ideals given by the hand transcriptions. They are biased by their training set – performance tends to be better on words that have many instances in the training corpus. Acoustic models may also compensate for pronunciation variation to some degree by smoothing out the phonetic classes, accepting variations within the canonical phonetic class estimates. It is important, therefore, to ascertain whether ASR systems perform worse in cases where there is a mismatch between the hand transcriptions and dictionary pronunciations.

For the WS96 system, we compared recognizer results in conditions where linguists determined that pronunciations were canonical versus conditions where alternative pronunciations were used by the speaker. In this study, we examined 439 sentences from the Switchboard development test set that were also phonetically transcribed. Each word in the test set transcriptions was annotated with whether it was correctly recognized, substituted, or deleted by the WS96 system, and whether the transcribers observed a canonical or alternative pronunciation, as defined by the Pronlex dictionary (i.e., the recognizer lexicon). Recognizer insertions were disregarded; although pronunciations certainly have an effect on insertions, it is diffcult to mark them as canonical or alternative pronunciations compared to the hand transcriptions, as the speakers did not actually utter the inserted words.

The WS96 system recognizes words correctly much more often when the linguists' transcription

[9] For the data set examined, the average word had 3.1 phones.

Table 4
Breakdown of word substitutions and deletions with WS96 Switchboard recognizer for canonical and alternative pronunciations

	Overall	Canonical pron.	Alternative pron.
% correct	57.4	65.0	53.9
% deleted	12.0	8.1	13.9
% substituted	30.5	26.1	32.2
# of words	4085	1337	2748

matches the dictionary pronunciation (Table 4). There is a large (70% relative) increase in the recognizer word deletion rate for words with alternative pronunciations, as well as a significant increase in recognizer substitutions. The fact that the recognizer accuracy for alternatively pronounced words is not very low, however, does indicate that there is some compensation for pronunciation variation by the acoustic model.

It is difficult to separate the effects of different factors on word error rates; for instance, a mispronounced word can result in a substitution, causing a language model error for the following word. Hence, some of the words labeled as having a canonical pronunciation may be identified incorrectly by the recognizer due to surrounding pronunciation errors; the extent of this phenomenon is difficult to characterize. Nevertheless, these numbers suggest that there is a real effect of pronunciations on word error. The numbers also show that solving "the pronunciation problem" will not necessarily solve the speech recognition problem, but will contribute towards reducing error rates.

5.2. Relationships between factors and recognizer error

Although we have seen that there is a relation between the pronunciation model and recognizer errors, it is not clear what the relationship is between recognizer errors and factors such as speaking rate, unigram probability, and trigram probability. For both corpora, we labeled every word in the development test set with the syllable rate, unigram probability, and trigram probability of the word. We then partitioned the words into histogram bins and determined the recognizer accuracy for each bin. The following series of graphs show how recognizer scores (y-axis) change as a function of each dynamic factor (x-axis). Included on each graph is the percentage of words that had canonical pronunciations and scores for words with or without canonical pronunciations, as marked by the transcribers.

In Fig. 8(a), we see that there is a 14% (absolute) drop in recognizer accuracy as the speaking rate moves from very slow to very fast speech. This is due mainly to the poorer performance on words pronounced non-canonically, which are more common in fast-speech conditions, as seen in Fig. 1. Note that for this test set the percentage of utterances in the fastest (>6 syllables/s) bin is non-trivial, containing 35% of the data; thus, there is a real and significant effect from fast speech for this set. One additional note: as in Section 5.1, these graphs do not include insertions. Since rate is calculated over an interpausal region, we can, however, calculate insertion rates for each speaking rate. Insertions decrease from 7.7% to 2.3% as the speaking rate increases; when this decrease in insertion rate is taken into account in the word error rate, the difference in errors between slow and fast speech is still roughly 9%.

In the case of language model probabilities (Fig. 8(b) and (c)), we do see that recognizer performance improves as words become more likely. This is not surprising, as both language models and acoustic models in the recognizer tend to favor more likely words during recognition. The trigram graph has a larger spread (from 30% to 69%) than the unigram (31% to 61%), probably because the recognizer (which utilizes a bigram grammar) takes into account more information than unigram probabilities. What is interesting here is that, even though the recognition rate increases as words become more likely, the percentage of words with canonical pronunciations decreases. [10] For higher proba-

[10] We are not certain why the probability of canonical pronunciations drops for low probability words. However, these words do tend to be longer on average, so a priori there is an increased chance of a single phone changing in a word. This class makes up 5% of the words in the test set.

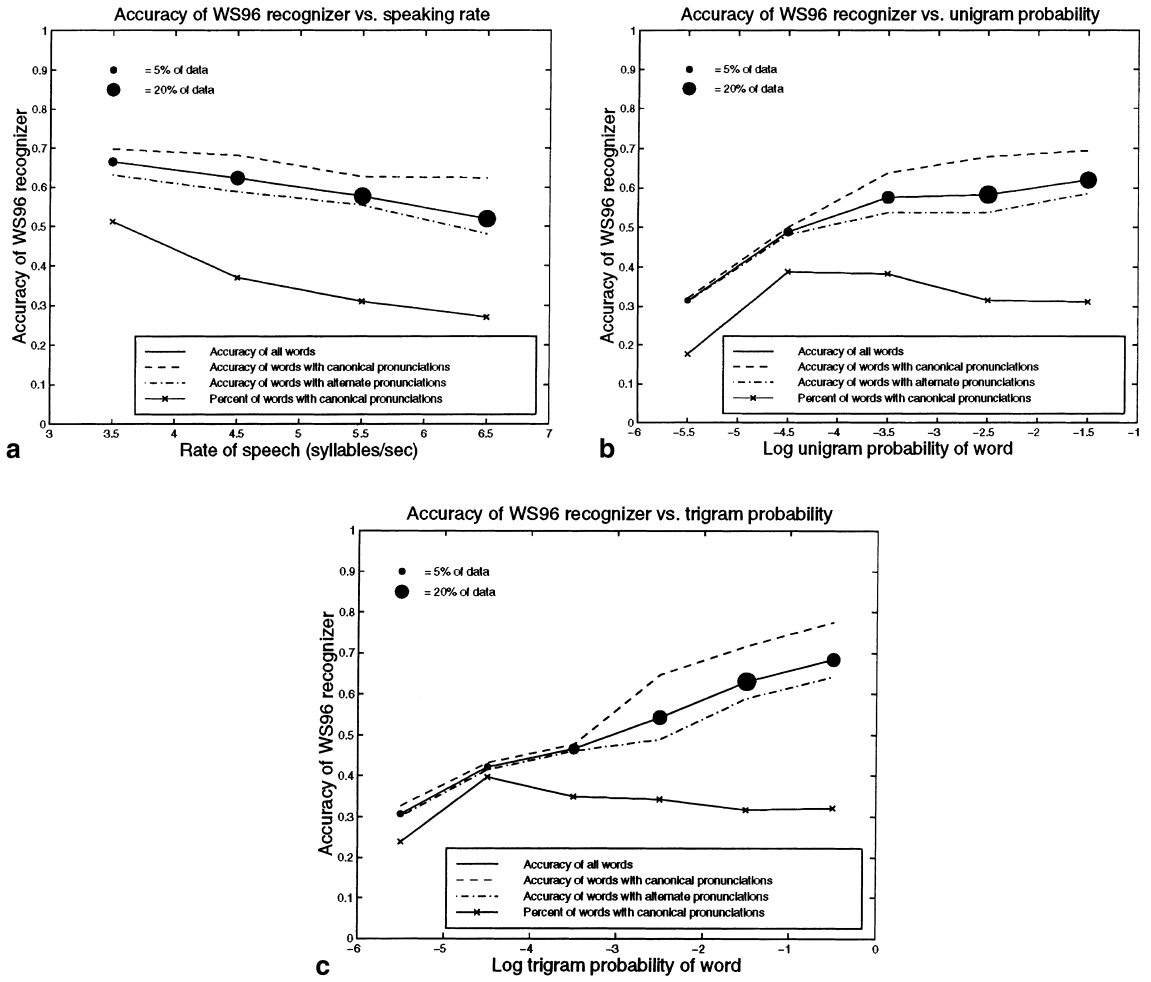

Fig. 8. Accuracy of WS96 Switchboard recognizer dependent on several factors. In these graphs, the solid line indicates the overall accuracy trend as each factor changes. The size of the dots indicate the proportion of data found in that particular histogram bin. The dashed and dot–dashed lines indicate recognizer scores when the hand transcription of the word did or didn't match the recognizer pronunciation, respectively. The solid line with crosses indicates the percentage of words that had canonical pronunciations for that histogram bin. (a) Speaking rate; (b) unigram probability; (c) trigram probability.

bility words (i.e., $\log_{10}$ (trigram) > -3), there is a gap in performance for canonical versus non-canonical pronunciations. On the other hand, for low probability words the language model in the recognizer dominates the error, and it does not matter as much whether the pronunciation was canonical or not. Therefore, it seems that there is a relationship between language model probabilities and pronunciations, although one must be careful to tease the effects apart from the influence of the language model itself in the recognizer.

5.3. Summary

In Switchboard, non-canonical pronunciation pervade the landscape; only 33% of words are canonically pronounced. The question is: are the acoustic models of the WS96 recognizer accommodating the pronunciation variation seen in this corpus? The models are certainly not doing a complete job, as words with non-canonical pronunciations have an 11% absolute increase in word error over canonically pronounced words.

Furthermore, we see that there is a correlation between word error and transcribed syllable rate. Faster speech goes hand-in-hand with increased errors, as has been observed in other corpora (Fisher, 1996b); much of this error can be attributed to the increase in pronunciation variation at fast rates. In the realm of word predictability, more likely words are recognized with better accuracy, as they are better modeled by the acoustic and language models. Pronunciation variation only affects the performance for highly probable words, whereas alternatively-pronounced words that are unlikely have roughly the same recognition accuracy as canonical baseforms.

6. Recognizer error analyses for transcription of Broadcast News

While knowing the correspondence between alternations in linguistic transcriptions and recognizer performance is important, many pronunciation modeling systems use automatic transcription methods to determine possible word pronunciations. Do the same patterns seen in the hand transcriptions of Switchboard carry over to an automatic learning paradigm? We have been working with such a pronunciation modeling scheme in the Broadcast News (BN) domain (Fosler-Lussier and Williams, 1999). We were therefore able to duplicate our Switchboard studies with this corpus using automatically determined transcriptions rather than hand alignments in the analysis. An added advantage to working with this corpus is the mixture of speaking styles; thus we could investigate the effects of our factors for both spontaneous and planned speech.

6.1. The corpus

The BN corpus (NIST, 1996) is a collection of speech from American radio and television news broadcasts, such as the National Public Radio program *All Things Considered* or *Nightline*, televised in the US on the ABC network. These shows comprise a whole range of speaking conditions, from planned speech in studio environments to spontaneous speech in noisy field conditions over telephone lines. The (possibly multi-sentence) segments are divided into seven different focus conditions representing different acoustic/speaking environments; we will be primarily investigating two main conditions that make up the majority of the data in the set: planned studio speech and spontaneous studio speech.

In the BN domain, we analyzed the results from the SPRACH hybrid neural network/HMM recognizer developed at Cambridge University, Sheffield University, and ICSI (Cook et al., 1999). This system combines a recurrent neural network trained on PLP features from Cambridge, a multilayer perceptron trained on modulation-filtered spectrogram features (Kingsbury, 1998) from ICSI, and HMM decoder technology from Sheffield.

In this investigation, we used an intermediate version of the evaluation recognizer described by Cook et al. (1999); the intermediate recognizer performed with roughly 20% more errors relative to the evaluation system. The acoustic models of the system described here were trained on 100 hours of BN speech. The lexicon of the recognizer used context-independent pronunciations from the Cambridge 1996 ABBOT recognizer (Cook et al., 1997); the trigram grammar was trained on 286 million words of text from transcriptions of broadcasts and newswire texts. We tested our system on a 173 segment subset of the 1997 BN DARPA evaluation test set, corresponding to roughly a half-hour of speech.

Unfortunately, we did not have the detailed phone-level transcriptions for BN that we had for Switchboard. In order to find an approximation to the phone transcription, we used our SPRACH BN recognizer in a phone-constrained decoding. The recognition of monophone models was constrained phonotactically using a phone-bigram grammar [11] in the phone recognizer. For each utterance, we generated an automatic phone transcription using the SPRACH BN recognizer and aligned the phone transcription to the word transcription. In step with the Switchboard anal-

[11] The phone-bigram grammar was trained using the phone transcription from a Viterbi alignment of the training set to the BN recognizer dictionary.

ysis, we also annotated test set words with whether the recognizer correctly identified them, substituted other words for them, or deleted them. In addition, the alignment was used to determine whether the word was pronounced canonically according to the recognizer's acoustic models.

While there is no guarantee that the phone transcription produced by the above procedure will match the decisions of human transcribers, [12] it does provide a clue to which acoustic models best match to the phonetic content of the waveform. Since the job of a pronunciation model is to facilitate matching between the acoustic models and word hypotheses in a recognizer, and since several researchers use phone recognition as a source for pronunciation alternatives, it is appropriate to investigate the effects of our dynamic variables on the automatic phonetic alignment.

6.2. Mismatch in pronunciations as a correlate of error

Using the BN database, we examined 173 (possibly multi-sentence) segments from the 1997 evaluation test set, which provided roughly the same number of words as the Switchboard test set. The difference between recognition rates for canonical versus non-canonical pronunciations is more marked for BN (Table 5); this is not unexpected, as the acoustic models of the recognizer determine whether a pronunciation is canonical, as opposed to the Switchboard analysis, which uses phonetic labelings provided by linguists. Both systems see a large increase in deletion rates for alternatively-pronounced words, but the increase in substitutions for these words is much greater for the SPRACH BN system – possibly due to the automatic phone transcription or to the larger overall error rate of the Switchboard system. It is also interesting to note that roughly the same proportion of words were judged to be pronounced canonically in each system, although the difference is significant ($p < 0.0001$) – 33% for Switchboard and 28% for BN.

Table 5
Breakdown of word substitutions and deletions with Hybrid BN recognizer for canonical and alternative pronunciations

	Overall	Canonical pron.	Alternative pron.
% correct	76.4	90.8	70.9
% deleted	4.7	1.4	6.0
% substituted	18.9	7.8	23.1
# of words	5840	1607	4233

6.3. Relationships between dynamic factors and recognizer error

When examining the language model graphs for BN (Fig. 9), the first difference from Switchboard that struck us was the increasing percentage of words having canonical pronunciations as words became more frequent (cf. Fig. 9(b) to Fig. 8(b)). This is probably an effect of the acoustic models: the recognizer is likely better at recognizing words found frequently in the training set, so the automatic phonetic transcription reflects this bias. Trigram scores are relatively flat, particularly for canonically pronounced words, although alternative pronunciation scores increase for the highest frequency and drop off for the lowest – the latter shows the influence of the language model in recognition.

The graph for unigram probabilities appears strange at first glance (Fig. 9(b)); instead of the smooth graph seen for Switchboard, the recognizer accuracy unexpectedly dips for words with a log unigram probability between −3 and −2. Further investigation revealed that the highest bin contained seven unique words [13] that are highly predictable from context. The second highest bin was dominated by a larger set of words that are less predictable, such as *is*, *this*, *it*, *who*, *well*, *years*, but are common enough that they would not normally receive extra emphasis in speech. The third bin held many "content" words that probably received stress in the sentence, such as

[12] We found via inspection of samples that the automatic phone recognizer usually produced intuitive transcriptions; however, conditions in which the acoustic models fare poorly, such as noisy speech, often degraded the phonetic transcript.

[13] These were *the*, *a*, *to*, *and*, *in*, *of* and *that*.

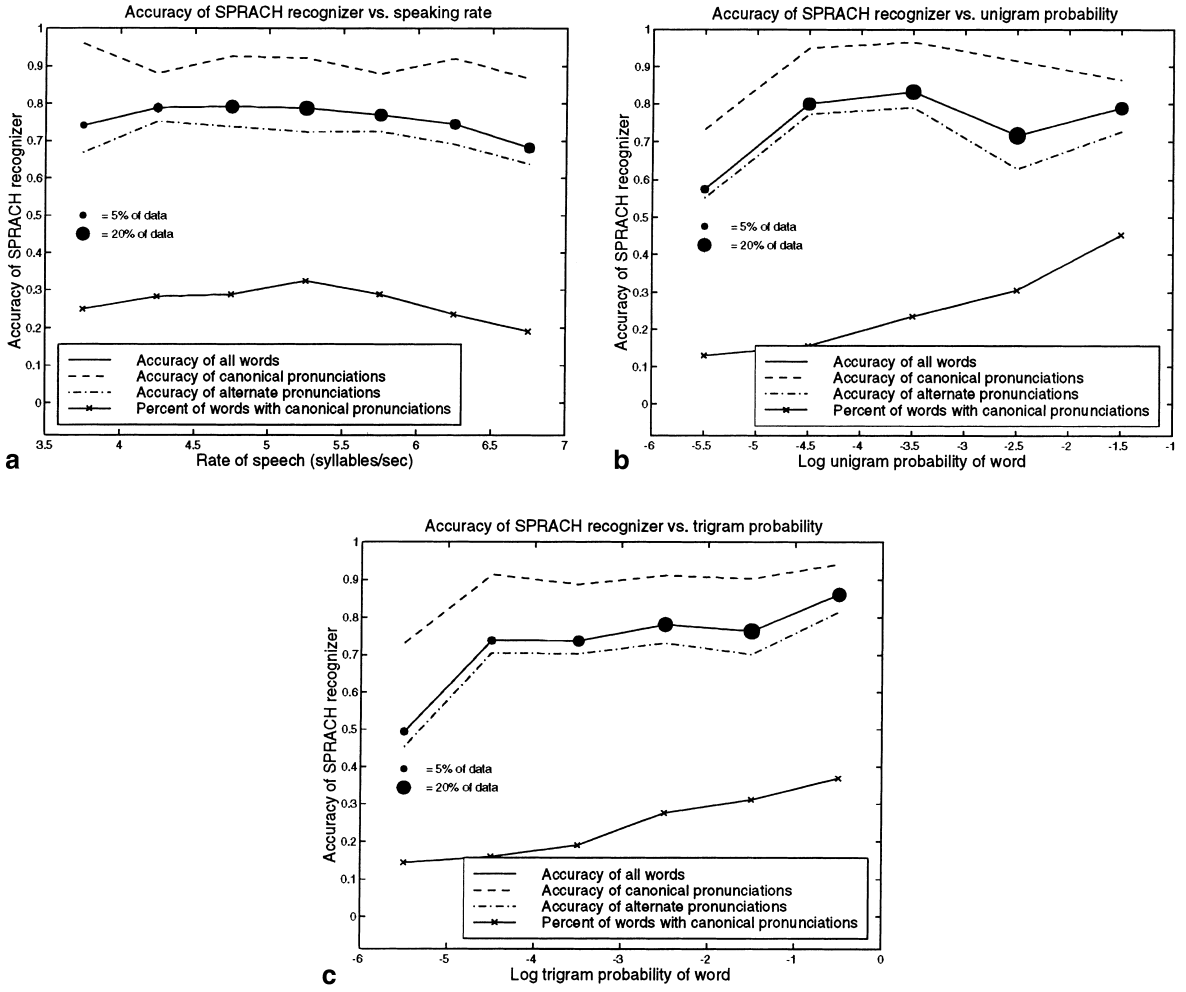

Fig. 9. Accuracy of SPRACH BN recognizer dependent on varying factors. (a) Speaking rate; (b) unigram probability; (c) trigram probability.

morning, *crime*, *campaign*, *economic*, and *American*. The third bin had more polysyllabic words (1.75 syllables/word average versus 1.19 syllables/word for the second bin); function words tend to be monosyllabic, while content words will range over a broader distribution. [14] It is likely that speakers emphasized these words more; anecdotally, stressed words are often clearer, and consequently are easier to automatically transcribe.

For speaking rate (Fig. 9(a)), we see that the percentage of words pronounced canonically peaks in the middle rates (5–5.5 syllables/s) and roughly tracks overall recognizer performance. There are several possible explanations for the shape of the curve: (1) the acoustic models are best when the speaking rate is roughly the mean, (2) the recognizer pronunciation model is geared toward

[14] Switchboard exhibits similar characteristics in its unigram grammar, although it is not as marked; for instance, 13 words occupy the most frequent unigram bin, and the number of syllables per word for the second bin (1.11) is still less than the for the third bin (1.51).

mean speaking rates, or (3) the speech is clearest in the mean speaking rates. From these data we cannot distinguish between these hypotheses, and it is likely that all are true to some extent. Recognizer performance suffers in both extremes, although not to the degree present in the Switchboard system.

When the data are separated out into planned and spontaneous conditions (Fig. 10), some of the differences in recognizer performance between these two speaking modes become apparent. For planned speech the difference between canonically and non-canonically pronounced words is much less than for spontaneous speech – in one histogram bin words with alternate pronunciations were recognized slightly more accurately than canonical words. Spontaneous speech is recognized much less consistently by this recognizer, and the performance gap between canonical/alternate pronunciations is very large, particularly for slow rates.

For an automatic phone-transcription system, the robustness of acoustic models has a serious impact on the pronunciation learning system. In Fig. 11, we have broken up the test set into different focus conditions and show the recognizer accuracies for each condition. The canonical pronunciation percentage tracks recognizer performance relatively well; when the acoustic/pronunciation models of the recognizer do well on the word level, the phone transcript from the acoustic model better matches the pronunciation model. Recognizer performance for noisy conditions is somewhat lower than the planned/spontaneous conditions; this is also reflected in the lower percentage of canonical pronunciations. While far from conclusive, this suggests that the quality of the acoustic models may be the cause of the poorer matching of the phonetic transcription to canonical models.

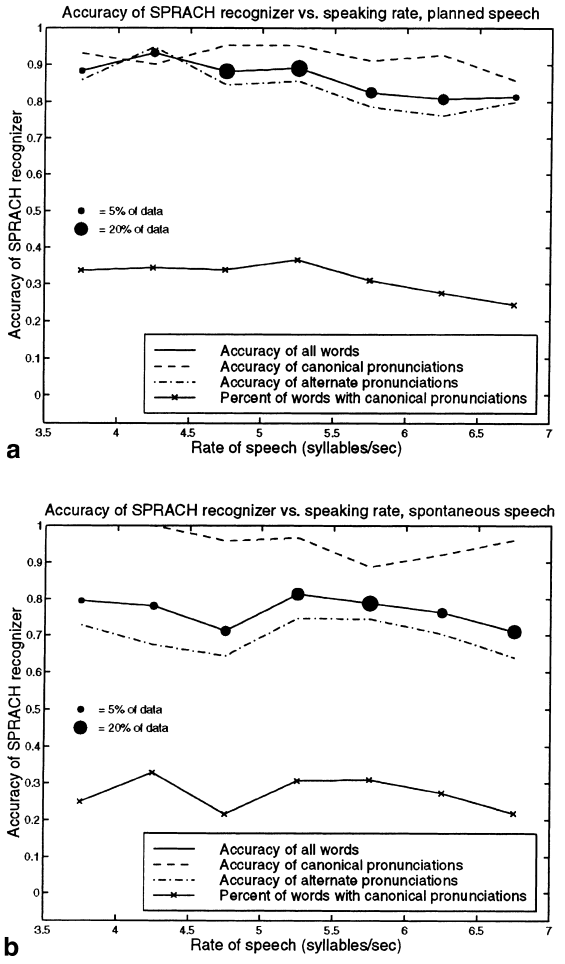

Fig. 10. Accuracy of Hybrid BN recognizer dependent on speaking rate for planned and spontaneous speech. (a) Planned speech; (b) spontaneous speech.

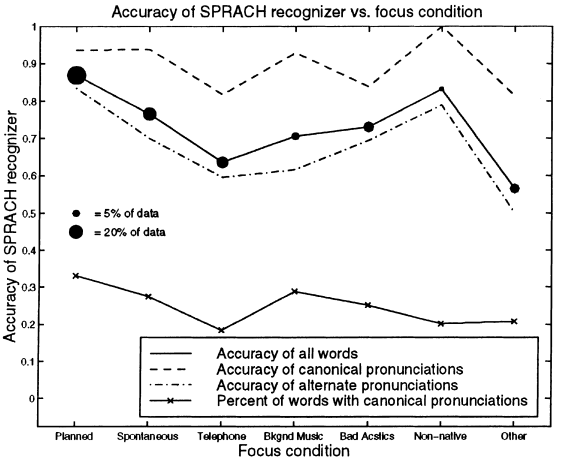

Fig. 11. Accuracy of Hybrid BN recognizer for different focus conditions.

6.4. Summary

Despite the dependency of the automatic phone transcription system on recognizer acoustic models, we found that there are distinct correlations in the BN database between pronunciation and recognizer error similar to those in the Switchboard database. Fast speaking rate again yields increased differences in the phonetic transcription, although, unlike Switchboard, distinct changes from the baseform dictionary were observed for slow speaking rates as well. The word predictability results are tied much more tightly to the automatic transcriptions: unlike in Switchboard, more likely words are transcribed canonically far more often. This built-in bias flattens out the accuracy curves, particularly in the trigram case.

Analyzing data from the BN test corpus allows us to compare spontaneous to planned speech. The drop in recognition accuracy for non-canonical pronunciations is much larger in the more casual speaking style. We also see, with the response of the recognizer to different acoustic conditions, that the robustness of the acoustic models has a distinct impact on an automatic pronunciation learning system.

7. Discussion and conclusions

We have argued that pronunciation models for ASR systems should be dynamic, taking into account contextual factors such as surrounding words, speaking rate, and word predictability. The investigations in this work have shown that pronunciations in the Switchboard corpus are strongly dependent on features corresponding to rate and language model probabilities. These pronunciation changes can be observed on the phone, syllable, and word level.

We were able to find relationships between pronunciations, rate, and word frequency. There is significant interplay between these features: syllabic distance from the canonical pronunciation is highest when word frequency and rate are both high, and trigram scores only become effective apart from unigram scores when the unigram is high. The complex interdependency between these variables makes sense from an information-theoretic viewpoint – since high-frequency words are more predictable, more variation is allowed in their production at various speaking rates, as the listener will be able to reconstruct what was said from context and few acoustic cues.

Furthermore, we have seen that these factors all affect recognizer performance on both the Switchboard and BN databases. Often, when there is more of a mismatch between the recognition dictionary and phonetic transcriptions, recognition performance decreases, corroborating the findings of McAllaster et al. (1998). The source of phonetic transcriptions has implications for finding these correlates; for automatic phonetic transcriptions, the quality of acoustic models seems to have a significant effect on the correspondence between the transcription and the recognizer dictionary.

While analyzing only one or two recognizers can certainly highlight the idiosyncrasies of those systems, the speech community has seen previously (Mirghafori et al., 1995; Siegler and Stern, 1995) that speaking rate affected the output of all systems in a 1993 ARPA evaluation on the Wall Street Journal corpus, so we have hopes that these studies may be applicable to more than just these systems.

We have integrated the lessons learned here into a dynamic syllable pronunciation model for rescoring *n*-best hypothesis lists within our ASR system. This model utilizes estimates of speaking rate, given by both the first pass recognition as well as our *mrate* measure, as well as unigram and trigram probabilities as features of word predictability. We also influence model construction with segmental context and duration features to estimate how syllables are pronounced in context. Initial experiments (Fosler-Lussier and Williams, 1999) suggest that this model is a good match for spontaneous speech; we currently achieve a 1.4% improvement (5% relative reduction in error) on the spontaneous studio (F1) focus condition of the BN corpus. We are continuing development of this promising paradigm.

Acknowledgements

We would like to thank Dan Jurafsky, Gethin Williams, Murat Saraclar, Sanjeev Khudanpur, Mitch Weintraub, Steve Greenberg, Brian Kingsbury, Jeff Bilmes, Dan Ellis, Michael Riley, Lori Lamel, three anonymous reviewers, and the ICSI Realization Group for helpful discussions of this work. We would also like to thank Bill Byrne for help with the WS96 recognizer, Gary Cook, Dan Ellis, Adam Janin, and Steve Renals for help with the Hybrid BN recognizer, and Steve Greenberg and the ICSI Switchboard Transcription Project for providing data. This work was supported by a grant from the Center for Language and Speech Processing at The Johns Hopkins University, NSF SGER grant IRI-9713346, and NSF grant IRI-9712579.

References

Bernstein, J., Baldwin, G., Cohen, M., Murveit, H., Weintraub, M., 1992. Phonological studies for speech recognition. In: DARPA Speech Recognition Workshop, Palo Alto, CA, pp. 41–48.

Bybee, J., 1996. The phonology of the lexicon: evidence from lexical diffusion. In: Barlow, M., Kemmer, S. (Eds.), Usage-based Models of Language.

Chen, F., 1990. Identification of contextual factors for pronounciation networks. In: IEEE ICASSP-90, pp. 753–756.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English, Harper and Row, New York, NY.

Cook, G., Kershaw, D., Christie, J., Robinson, A., 1997. Transcription of broadcast television and radio news: The 1996 ABBOT system. In: DARPA Speech Recognition Workshop, Chantilly, VA.

Cook, G., Christie, J., Ellis, D., Fosler-Lussier, E., Gotoh, Y., Kingsbury, B., Morgan, N., Renals, S., Robinson, T., Williams, G., 1999. The SPRACH system for the transcription of broadcast news. In: DARPA Broadcast News Workshop, Herndon, VA.

Finke, M., Waibel, A., 1997a. Flexible transcription alignment. In: 1997 IEEE Workshop on Automatic Speech Recognition and Understanding, Santa Barabara, CA, pp. 34–40.

Finke, M., Waibel, A., 1997b. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In: Eurospeech-97.

Fisher, W., 1996a. Factors affecting recognition error rate. In: DARPA Speech Recognition Workshop, Chantilly, VA.

Fisher, W., 1996b. The tsylb2 Program: Algorithm Description. NIST. Part of the tsylb2-1.1 software package.

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, VA.

Ganong, W., 1980. Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Performance and Perception 6, 110–125.

Gauvain, J.L., Adda, G., Lamel, L., Adda-Decker, M., 1997. Transcribing broadcast news: The LIMSI Nov96 Hub4 system. In: DARPA Speech Recognition Workshop, Chantilly, VA.

Greenberg, S., 1997. WS96 project report: The Switchboard transcription project. In: Jelinek, F. (Ed.), 1996 LVCSR Summer Research Workshop Technical Reports, Center for Language and Speech Processing, Johns Hopkins University, Chapter 6.

Greenberg, S., 1998. Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation. In: ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Kerkrade, The Netherlands, pp. 47–56.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., Raymond, W., 1998. Reduction of English function words in Switchboard. In: ICSLP-98, Sydney, Australia.

Kahn, D., 1980. Syllable-Based Generalizations in English Phonology. Garland, New York.

Kingsbury, B.E.D., 1998. Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments. Ph.D. thesis, University of California, Berkeley, CA.

Kitazawa, S., Ichikawa, H., Kobayashi, S., Nishinuma, Y., 1997. Extraction and representation of rhythmic components of spontaneous speech. In: Eurospeech-97, Rhodes, Greece, pp. 641–644.

Linguistic Data Consortium (LDC), 1996. The PRONLEX pronunciation dictionary. Available from the LDC, ldc@unagi.cis.upenn.edu. Part of the COMLEX distribution.

McAllaster, D., Gillick, L., Scattone, F., Newman, M., 1998. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. In: ICSLP-98, Sydney, Australia, pp. 1847–1850.

Miller, J., Grosjean, F., 1981. How the components of speaking rate influence perception of phonetic segments. Journal of Experimental Psychology: Human Performance and Perception 7 (1), 208–215.

Mirghafori, N., Fosler, E., Morgan, N., 1995. Fast speakers in large vocabulary continuous speech recognition: Analysis & antidotes. In: Eurospeech-95.

Mirghafori, N., Fosler, E., Morgan, N., 1996. Towards robustness to fast speech in ASR. In: ICASSP-96, Atlanta, Georgia, pp. I335–338.

Morgan, N., Fosler-Lussier, E., 1998. Combining multiple estimators of speaking rate. In: IEEE ICASSP-98, Seattle, WA.

Morgan, N., Fosler, E., Mirghafori, N., 1997. Speech recognition using on-line estimation of speaking rate. In: Eurospeech-97.

NIST, 1992. Switchboard corpus: Recorded telephone conversations. National Institute of Standards and Technology Speech Disc 9-1 to 9-25.

NIST, 1996. 1996 broadcast news speech corpus. CSR-V, Hub 4, Produced by the Lingustic Data Consortium.

Ostendorf, M., Byrne, B., Bacchiani, M., Finke, M., Gunawardana, A., Ross, K., Roweis, S., Shriberg, E., Talkin, D., Waibel, A., Wheatley, B., Zeppenfeld, T., 1997. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In: Jelinek, F. (Ed.), 1996 LVCSR Summer Research Workshop Technical Reports, Center for Language and Speech Processing, Johns Hopkins University, Chapter 4.

Pallett, D.S., Fiscus, J.G., Fisher, W.M., Garofolo, J.S., Lund, B. A., Przybocki, M.A., 1994. 1993 WSJ-CSR benchmark test results. In: ARPA Spoken Language Systems Technology Workshop, Princeton, NJ.

Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern, R., Thayer, E., 1997. The 1996 hub-4 sphinx-3 system. In: DARPA Speech Recognition Workshop, Chantilly, VA.

Riley, M., 1991. A statistical model for generating pronunciation networks. In: IEEE ICASSP-91, pp. 737–740.

Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., 1998. Stochastic pronunciation modelling from hand-labelled phonetic corpora. In: ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition, Kerkrade, The Netherlands, pp. 109–116.

Saraclar, M., 1997. Automatic learning of a model for word pronunciations: Status report. In: Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation, Baltimore, MD.

Siegler, M.A., Stern, R.M., 1995. On the effects of speech rate in large vocabulary speech recognition systems. In: IEEE ICASSP-95.

Sloboda, T., Waibel, A., 1996. Dictionary learning for spontaneous speech recognition. In: ICSLP-96.

Summerfield, Q., 1981. Articulatory rate and perceptual constancy in phonetic perception. Journal of Experimental Psychology: Human Performance and Perception 7, 1074–1095.

Tajchman, G., Fosler, E., Jurafsky, D., 1995. Building multiple pronunciation models for novel words using exploratory computational phonology. In: Eurospeech-95, Madrid, Spain.

Verhasselt, J.P., Martens, J.-P., 1996. A fast and reliable rate of speech detector. In: ICSLP-96, Philadelphia, PA, pp. 2258–2261.

Weintraub, M., Fosler, E., Galles, C., Kao, Y.-H., Khudanpur, S., Saraclar, M., Wegmann, S., 1997. WS96 project report: Automatic learning of word pronunciation from data. In: Jelinek, F. (Ed.), 1996 LVCSR Summer Research Workshop Technical Reports, Center for Language and Speech Processing, Johns Hopkins University, Chapter 3.

Withgott, M.M., Chen, F.R., 1993. Computational Models of American Speech, Center for the Study of Language and Information, Stanford, CA.

Young, S.J., Odell, J.J., Woodland, P.C., 1994. Tree-based state tying for high accuracy acoustic modelling. In: IEEE ICASSP-94, pp. 307–312.