Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition

Matthew A. Siegler

Submitted to the Department of Electrical and Computer Engineering in Partial Fulfillment of the Requirements for the Degree of Master of Science at

Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

December 1995

Abstract

This report describes a series of experiments that measure speech rate and that attempt to improve speech recognition accuracy for rapidly-spoken speech. Descriptions of several measures of speech rate are presented, with their advantages and disadvanatges. Speech recognition results obtained using several compensation methods are compared to identify methods by which compensation for the effects of fast speech may yield the greatest improvement in recognition accuracy.

Very simple measures of speech rate such as the word rate or phone rate are found to be unsuitable for detection of both long-term and short-term speech rate since they are sensitive to the lexical content of speech. In contrast, the phone duration percentile, a comparison of measured versus expected phone duration, is shown to be robust with respect to lexical content and consistent with previous findings about the statistics of long-term and short-term speech rate. Using this metric, speakers with a speech rate in the top 30% are found to produce a 50 to 150% increase in word error rate.

The compensation techniques explored contain modifications to five components of the recognition system: the models of the acoustical characteristics of speech sounds, the models of the HMM state-transition probabilities, the pronunciations of words in the dictionary, the weight with which acoustic and linguistic evidence are combined, and the base phone set. Optimizing the language weight reduced the word error rate of fast speech by 10.3% relative to baseline performance. Adapting the state-transition probabilities to fast speech reduced the word error rate for fast speech by 2.6%. Using one of the modified pronunciation dictionaries reduced the word error rate of fast speech by 2.6%. The other techniques yielded little or no reduction in the word error rate.

Acknowledgments

I would like to thank the following people for making this research and report possible: Dr. Rich Stern, my advisor and friend for his guidance, support, and wisdom of which I am very grateful. Dr. Raj Reddy, for sponsoring my research position at Carnegie Mellon University. Robert Weide, who provided a much needed linguist's opinion for this work. Pedro Moreno and Bhiksha Raj, for their expert insight in the fields of acoustic compensation. Uday Jain, for his opinions on countless questions regarding the construction and interpretation of many experiments. Eric Thayer, Ravi Mosur, and Paul Placeway for their continuing support of the SPHINX-II training and decoding software. And finally, my future wife Erika, for her compassion in the midst of chaos in the composition of this report.

Table of Contents

Chapter 1: Introduction

As the performance of speaker independent (SI) continuous speech recognition (CSR) has improved over the last decade, increasing attention has been given to the degradation of performance for specific speakers. In recent years, fast speech rate has been identified as a speaker characteristic that can significantly impair recognition accuracy.

Previous studies of speech generation and perception have shown that the phonological effects of varying speech rate are complex and sometimes subjective. However, several principles have emerged from this research which can be applied to CSR.

In contrast, little attempt has been made to identify a suitable metric for speech rate in the context of CSR. Such a metric would be robust to the phonetic and lexical content of the speech and be able to resolve changes in speech rate at the word level.

Even with knowledge of speech rate, compensation techniques for the effects of variations in speech rate have been limited so far to the use of phone duration modelling. Largely, these compensation techniques are unsuccessful.

The purpose of this work is to identify a measure of, to evaluate the effects of, and to compensate for speech rate with the intention of improving the overall performance of a specific speech recognition system.

The outline of this report is as follows:

- Chapter 2 is a survey of previous work in the area of speech rate.
- Chapter 3 explains components of the SPHINX-II CSR system relevant to this research.
- In Chapter 4, a robust measure of speech rate is identified and evaluated.
- Chapter 5 describes several speech rate compensation techniques and their effectiveness.
- Chapter 6 summarizes this research and provides suggestions for future work.

Chapter 2: Previous Studies in Speech Rate

Speech rate has been identified in the past as an important phenomenon in the generation and recognition of speech by humans and computers. This chapter is a discussion of previous research into the causes and effects of speech rate in both a phonological sense and within the context of continuous speech recognition.

2.1 Phonological Studies of Speech Rate

Most previous studies have been performed with varying scopes of measurement: typically at the sentence level for long-term effects, and at the word level for short-term effects. Because there is little agreement on a precise definition of short-term speech rate, most phonological research has concentrated on the effects of long-term speech rate as defined in very simple terms. To date, there has been no attempt to develop a speech rate metric that is consistent with both long and short-term observations.

2.1.1 Long-Term Speech Rate

Researchers have found that the long-term statistics of speech rate are primarily dependent on two properties: the style of the speaker and the nature of the text. Generally, the mean speech rate is a function of the speaker and the variance is a function of the cognitive load associated with the text [23]. Cognitive load is defined loosely here as the level of effort and creativity required to select the words to speak. For example, in a fully spontaneous dialogue between two individuals, the cognitive load is very high and thus the speech rate varies greatly from sentence to sentence. In comparison, reading sequences of digits from prepared texts has a very low cognitive load. For fluent read speech such as the kind found in this work, the cognitive load is relatively low and therefore the variance of long-term speech rate tends to be small [23].

Transformations of the acoustics of speech when the speech rate increases have been of particular interest. A few regular patterns emerge when speech rate increases and some have classified these as transformations that map fast speech phonemes onto normal speech phonemes [31].

There are many studies of how the durations of phones changes when the rate of speech increases. Overall, vowels show more changes than consonants and so there has been much focus on them [27][16][11]. How much the presence of fast speech will reduce the duration of a vowel depends on he kind of vowel [27][16], the adjacent consonants [27][28], and the kind of word [36]. The combination of these effects is neither simple nor linear [7]. Unstressed vowels are especially susceptible to the shortening effects of fast speech. Indeed, schwas are frequently deleted entirely in fast speech [27][16]. Content words seem to be more resistant to the effects of fast speech as their vowels are not affected nearly as much [36].

As speech rate increases, the acoustic variety associated with any particular phone increases [35][6][24][16]. By the same token, the difference between phone acoustics decreases and the confusion between phones increase. Humans have trouble identifying fast phones when they are inserted into normal speech, demonstrating the importance of speech rate adaptation [20][6].

2.1.2 Short-Term Speech Rate

O'Shaughnessy (1995) has correlated the short-term statistics of speech rate with a large variety of phenomenon such as the presence of proper names or the introduction of a new term [23]. However, the phonological effects have not yet been established for short-term speech rate.

2.2 CSR Studies

Pallett *et al.* (1994) showed clearly that the presence of fast speech causes a degradation in performance of automatic recognition [25]. Most researchers that have attempted to compensate for fast speech use phone duration modeling either during recognition [34][21][3] or in post-processing [22][2]. Their methods do not appear to improve recognition significantly. However, they all agree that the state-transition probabilities inherent in acoustic modeling using hidden markov models are problematic with fast speech. In addition, Anastasakos, *et. al*. (1995) found that large fluctuations in the short-term speech rate are correlated with high recognition errors [2].

Chapter 3: The SPHINX-II Speech Recognition System

In this project, the SPHINX-II recognition system was used to perform all experiments. Since it is a very complex system, only those elements that were manipulated for compensation are discussed here. SPHINX-II uses state-tied hidden Markov models. Good tutorials on hidden Markov models for speech recognition have been written by either Huang *et al*. [14] or Rabiner and Juang [29].

3.1 Acoustic Models

The basic unit of speech in SPHINX-II is termed the **base phone**. The 50 base phones that represent speech sounds are shown in Figure 3-1. An additional 5 represent non-speech filler sounds such as silences and extraneous noises. Each base phone has a **context-independent** model as well as multiple **context-dependent** models, depending on whether information about left or right contexts is used. In either case, each phone has a 5-state forward-only hidden Markov model including transitions from states 1 to 3 and 3 to 5. The upper part of Figure 3-2 shows a context-dependent phone and its components.

AA	DH	K	T
AE	DX	KD	TD
AH	EH	L	TS
AO	ER	M	UH
AW	EY	N	UW
AX	F	NG	V
AXR	G	OW	W
AY	GD	OY	Y
B	HH	P	Z
BD	IH	PD	ZH
CH	IX	R	
D	IY	S	
DD	JH	SH	

Figure 3-1 Base phones used in the SPHINX-II system

In feature extraction the 16 kHz sample rate input signal is split into 20 ms windows, called **frames**, occurring every 10 ms, for a frame rate of 100 frames per second. Each frame is processed to obtain 12 mel-frequency warped cepstral coefficients [13]. Then, four independent sets of features are extracted from the cepstral coefficients: the cepstra, the first-order difference cepstra, the second-order difference cepstra, and the power. These form a total of 51 parameters. The acoustic space of each feature set is modelled as a mixture gaussian with four components. Details on this process can be found in [13].

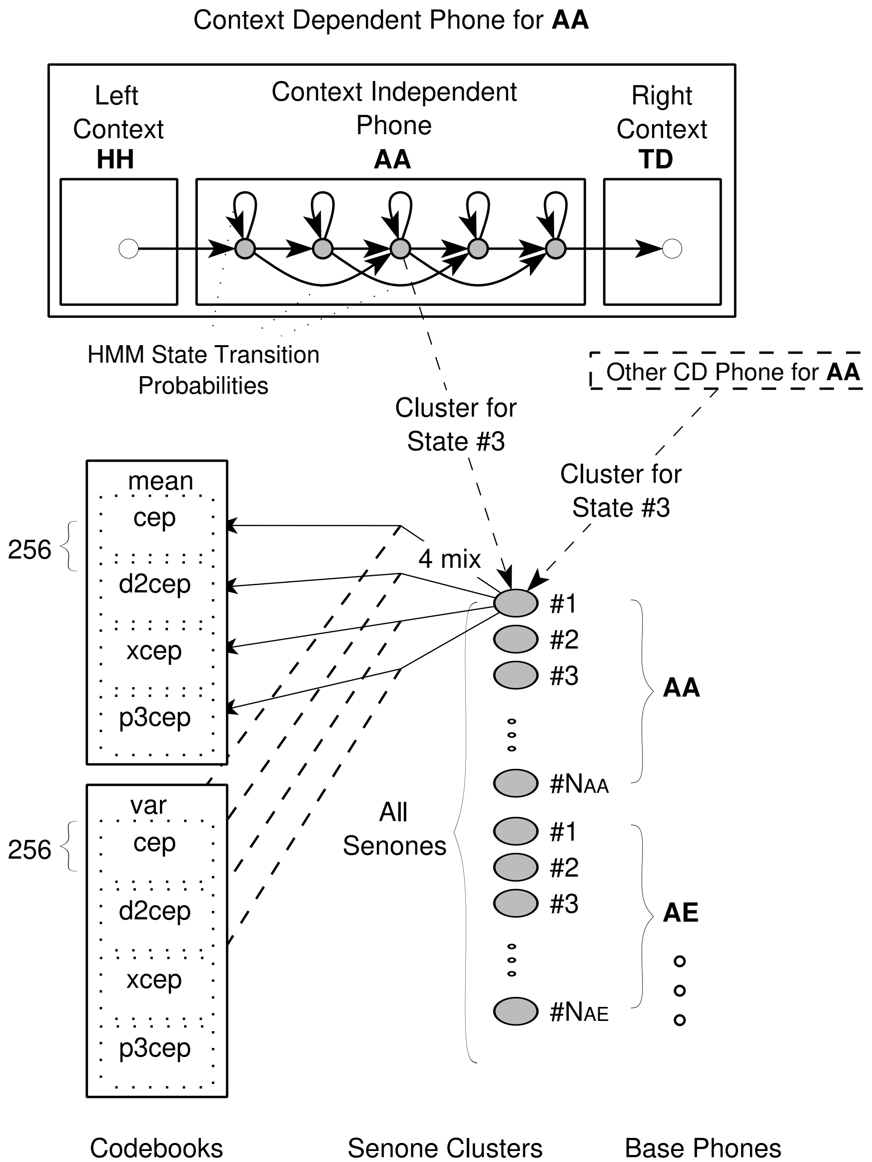

Figure 3-2 Acoustic modeling elements in SPHINX-II. The upper section shows the structure of a phone. The lower section shows the way states are clustered, and how parameters are tied together. See text for details.

The way these components, the **means**, **variances** and **mixture weights**, are allotted to different speech events is the most critical part of the acoustic modeling in SPHINX-II. In a method called **parameter-tying** parameters are shared between different kinds of acoustic events to reduce the total number of variables. For example, only 256 distinct mean and variance locations are kept in the **codebook**, shared by all phones.

Each state of every context-dependent model, and there may be more than 100,000, must be assigned four mixtures weights and the indicies of four means and variances from the codebook. Because of the high dimensionality and the very large number of possible context-dependent phones, the states are shared across context-dependent phones. These shared states in SPHINX-II are called **senones**. A large effort is placed on the manner of state-tying across different contexts for the same context-independent phone. This means that the acoustic model for a particular state of a context-dependent model will be exactly the same as that for the same state of a different context-dependent model. The lower part of Figure 3-2 illustrates the mixture gaussians and their assignments to senones across many different base phones. An important note is that the states cannot be shared between different context-independent phones. For example, state 3 of AA in the sequence HH AA TD cannot be shared with state 3 of AE in the sequence HH AE TD

3.2 Lexical Models

Words are modeled in SPHINX-II with a **pronunciation dictionary**. Each entry in the dictionary is a word and a sequence of base phones selected by a linguist to represent the word. To accommodate different accents each word may have a number of alternative pronunciations in addition to the base form. Figure 3-3 shows an excerpt from a dictionary.

3.3 Language Model

SPHINX-II uses a statistical trigram language model to compute the probability of sequences of words that are hypothesized from the acoustic search [30]. Each word sequence in the hypothesis is broken down into overlapping groups of three words and probabilities are computed from the model. The model itself is built using many millions of words of training material from sources which are similar to the expected application of the recognition system. For example, in the experiments in this work the target application is the dictation of news articles. To train the language model, many articles from the Wall Street Journal were used.

Figure 3-4 is an excerpt of the language model similar to the one used for the experiments in this work. Each line contains a log probability and sequence of three words. The log probability is the conditional likelihood of those three words given that exactly three words actually occur.

APPAREL	AX P AE R AX L
APPAREL(2)	AX P EH R AX L
APPARENT	AX P EH R AX N TD
APPARENTLY	AX P EH R AX N TD L IY
APPEAL	AX P IY L
APPEALED	AX P IY L DD
APPEALING	AX P IY L IX NG
APPEAR	AX P IH R
APPEARED	AX P IH R DD
APPEARS	AX P IH R Z
APPLE	AE P AX L
APPLICATIONS	AE P L AX K EY SH AX N Z
APPLIED	AX P L AY DD
APPOINTED	AX P OY N T AX DD
APPOINTED(2)	AX P OY N T IX DD
APPRECIABLE	AX P R IY SH AX B AX L
APPRECIATE	AX P R IY SH IY EY TD
APPRECIATION	AX P R IY SH IY EY SH AX N
APPROACH	AX P R OW CH
APPROACHED	AX P R OW CH TD
APPROPRIATE	AX P R OW P R IY AX TD
APPROPRIATE(2)	AX P R OW P R IY EY TD

Figure 3-3 Excerpt of a pronunciation dictionary.

```
-1.0792 BUT IN INTERVIEWING
-1.6812 BUT IT ALSO
-1.6812 BUT IT DIDN'T
-1.6812 BUT IT DOESN'T
-1.6812 BUT IT WAS
-1.0792 BUT MANY OF
-1.0792 BUT OVERALL THE
-1.0792 BUT SHE NEVER
-1.0792 BUT THAT'S ONE
-1.6812 BUT THE CONSERVATION
-1.6812 BUT THE EARTHQUAKE
-1.6812 BUT THE LATEST
-1.6812 BUT THE OTHER
```

Figure 3-4 Excerpt of a language model.

3.4 Combing the Models in the Decoder

Since the acoustic, lexical, and language models are trained and tested separately and not jointly, the manner in which the likelihood or **scores** from each model are combined is critical to good recognition performance. To start with, SPHINX-II performs recognition in several passes, depending on the configuration at runtime. For the experiments in this work, the most comprehensive search strategy is performed where the acoustic, lexical, and language models are all simultaneously evaluated. The result of the search is a set of 100 hypotheses and the acoustic and language model scores for each hypothesis. In a process called **N-best rescoring** the scores are weighted again and added according to the **language weight** and then ranked by likelihood. It is at this stage where any other evidence could be used to rescore the hypothesis. The most likely sentence is then emitted as the decoder hypothesis. These steps are illustrated in Figure 3-4.

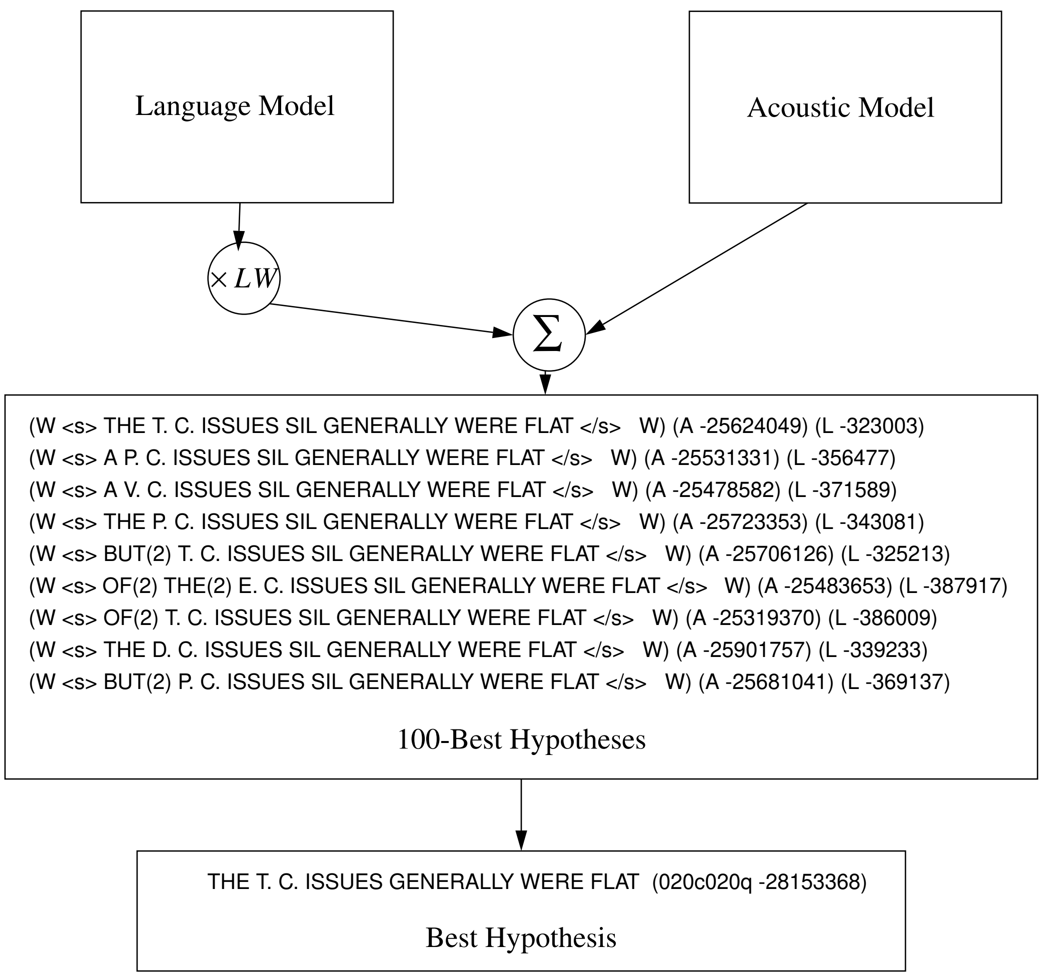

Figure 3-5 Combining evidence from multiple modeling sources.

Chapter 4: Measures and Effects of Speech Rate

As discussed in Chapter 2, there is strong evidence that the performance of continuous speech recognition (CSR) systems degrades for speakers with abnormal speech rates [25][32]. Still, little attempt has been made to identify a metric for speech rate that is text-independent. The goal of this chapter is to develop such a speech rate metric, examine the performance of automatic recognition of speech at different rates, and to evaluate different techniques for classifying and estimating the speech rate for any particular speaker.

This chapter is divided as follows:

- Definition of three speech metrics
- Effects of each speech metric on recognition performance
- Robustness of each speech metric
- Classification strategies for clustering with respect to speech rate
- Estimation of speech rate or the class of speech rate

4.1 Speech Rate Metrics

Because it is a stylistic property of a given speaker, fluent speech is characterized by a fairly constant speech rate. However, it would also be beneficial to able to indicate local changes in speech rate as well because these changes occur at major semantic and syntactic boundaries [23].

There are a variety of ways to analyze the global and local speech rate, chiefly varying in the amount of detail used to analyze the speech. More detailed analysis in this sense means the use of more *a-priori* knowledge in measurement. In the sections up to and including 4.3 the assumption is that the transcription is known, and that phone-level segmentation can be obtained through Viterbi alignment [14].

4.1.1 Word Rate

Pallett *et al.* [25] used the **word rate** measure to compute the speech rate of utterances from the WSJ1 corpus. In their experiments, word rate was calculated by dividing the number words in the transcript by the total length of the utterance in minutes. Figure 4-1 is a plot of the word rate for the WSJ01 corpus.

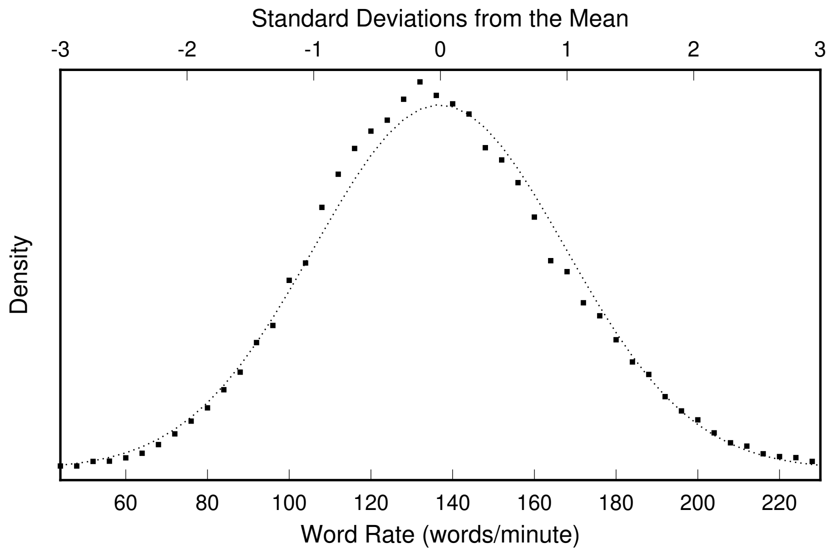

Figure 4-1 Histogram of the word rate for the WSJ1 corpus. The dotted line is the best fit Gaussian. The mean value is 137 and the standard deviation is 31.

Campbell(1988) pointed out that the word rate is unsatisfactory "because of unpredictability in the structure and length of a word, which may be monosyllabic or polysyllabic, and because of the indeterminacy of any pause durations between words" [4]. In light of this, a more precise measure of speech rate must characterize the rate of information using a much smaller unit than the word.

4.1.2 Phone Rate

For each phone in an utterance the instantaneous phone rate is defined as the inverse of the phone duration. It is very difficult to automatically determine the exact boundaries of the phones within an utterances, and so the average phone rate over a number of phones yields a smaller estimation error. This is computed

$$\frac{N}{\sum_{i = 1 \ldots N} d_i} \tag{4.1}$$

where N is the number of phones and d_i is the duration of phone i.

For SPHINX-II, phone segmentation errors are on the order of 2 frames or approximately 20ms. If the average phone duration is 6 frames, the error in the estimate of the phone rate of a single phone will be 30%.

However, when 3 phones are averaged, the error drops to 10%. This last measure was used in the analysis here.

The mean phone rate over the entire utterance, excluding silence regions, was used to classify an utterance. Since the phone rate is not directly related to word length, it is unaffected by the preponderance of long or short words in a given utterance. Figure 4-2 shows the distribution of sentence-averaged phone rate in the training set.

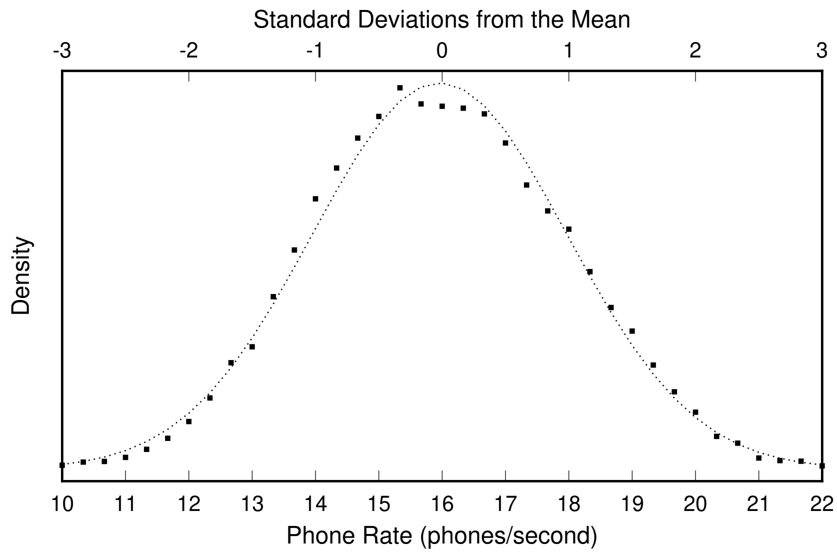

Figure 4-2 Histogram of the phone rate for the WSJ1 corpus. The dotted line is the best fit Gaussian. The mean value is 16.0 and the standard deviation is 2.0.

4.1.3 Phone Duration Percentile

One problem with the phone rate is that the inherent duration statistics of a phone varies with its identity as shown in Figure 4-3. For example, a sentence containing many unstressed vowels and unvoiced stops will have a much higher phone rate than a sentence containing many stressed vowels and fricatives.

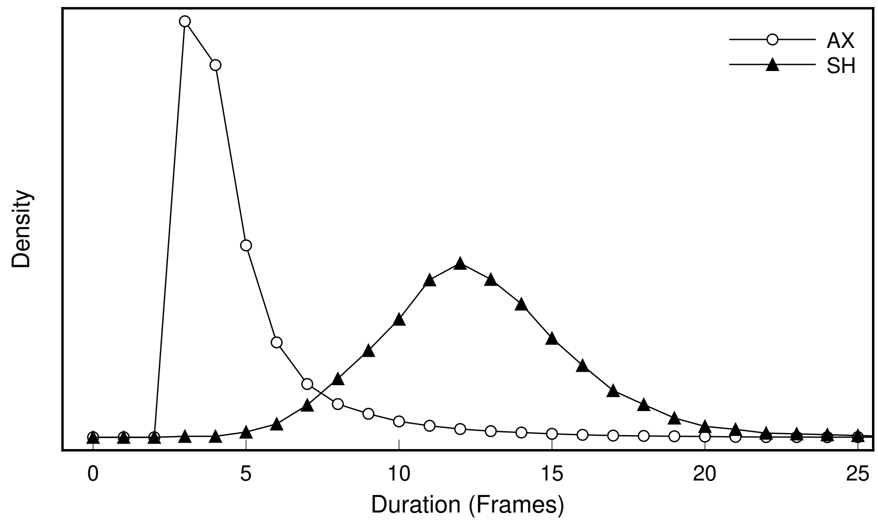

Figure 4-3 Histograms for the duration of two different phones.

One common alternative is to base the rate of speech on comparisons of observed phone duration with predicted phone duration [2][3][21][34]. The duration of each phone (or phone in context) can then be modelled with a probability density function (PDF) estimated from the training data. The Log-Normal [21], Gamma [7][3] and Poisson distributions have all been used in duration modelling, but it is important to note that the duration statistics for a particular context independent unit are not necessarily unimodal let alone Gaussian [36]. Many vowels exhibit at least bimodal distributions due to the impact of stress on vowel duration [36][11]. In addition, consonants can also have bimodal distributions due to lexical stress [37].

Ideally, duration statistics should be computed with left and right context and word-boundary information since each has an effect on phone duration [36][37][28]. However, the lack of training data prohibits such detailed modelling. Some have used triphone clustering methods to allow CI phones in different contexts [21] to share duration statistics based on regression tree clustering methods [15], others have used context back-off schemes [2].

Although the duration PDF provides the probability of occurrence for observed phone durations, it does not indicate the relative duration in comparison to previously observed data or the relative rate of speech.

Some have used the ratio of the measured phone rate to the mean phone rate [2][21][34], which yields a modal distribution. For this work, the cumulative distribution function of the observed phone duration or **phone duration percentile** is used instead. The percentile yields an interval bound measure of the rate of speech for any number of phones in sequence.

4.1.3.1 Derivation of the Rate of Speech

In computing the relative rate of speech for a single phone, some assumptions need to be made about the duration distributions to simplify the measurement. It is first assumed that the duration of a phone is independent of its context.

$$Pr\langle d_i | p_1, p_2, \ldots, p_I \rangle \approx Pr\langle d_i | p_i \rangle \tag{4.2}$$

where d_i is the duration of phone i and p_i is the identity of phone i from a total sequence of I phones.

For the reasons discussed above, the cumulative distribution function is used instead. Cumulative distributions are taken from the right instead of the left so that shorter durations (and hence faster rates) yield higher numbers:

$$\mathrm{ROS}_i \equiv Pr\langle d > d_i | p_i \rangle \tag{4.3}$$

The probability of the duration of a phone given its identity was calculated using automatic segmentation. Histograms were collected instead of using a prior distribution.

4.1.3.2 Smoothing Windows

For duration modelling, Anastasakos *et al.* [2] have used a measure of rate averaged over multiple units so that segmentation errors are smoothed out over multiple segments. If the duration of a phone d_i cannot be measured accurately, as may be the case due to segmentation errors, the problem must be reformulated. What can be measured more accurately (to a given certainty) is the total duration of a sequence of N phones. The phone duration probability measure then can be viewed as dependent on the previous N phones.

$$Pr\langle d_i, d_{i+1}, \ldots, d_{i+N-1} | p_i, p_{i+1}, \ldots, p_{i+N-1} \rangle \approx$$
$$Pr\langle d_i + d_{i+1} + \ldots + d_{i+N-1} | p_i, p_{i+1}, \ldots, p_{i+N-1} \rangle \tag{4.4}$$

Assuming that duration probabilities are context independent this can be reformulated as

$$Pr\langle d_i, d_{i+1}, \ldots, d_{i+N-1} | p_i, p_{i+1}, \ldots, p_{i+N-1} \rangle \approx$$
$$\{ Pr\langle d|p_i \rangle \bullet Pr\langle d|p_{i+1} \rangle \bullet \ldots \bullet Pr\langle d|p_{i+N-1} \rangle \} \Big|_{d = d_i + d_{i+1} + \ldots + d_{i+N-1}} \tag{4.5}$$

Taking the cumulative distribution ROS for a phone becomes

$$\mathrm{ROS}_N \equiv \{ Pr\langle d|p_i \rangle \bullet Pr\langle d|p_{i+1} \rangle \bullet \ldots \bullet Pr\langle d|p_{i+N-1} \rangle \} \Big|_{d > d_i + d_{i+1} + \ldots + d_{i+N-1}} \tag{4.6}$$

Examining the ROS statistics for any specific speaker with a window N<6, the distribution is either uniform on [0,1], or tail-heavy. Figure 4-4 shows the distributions for the phone duration percentile for three speakers.

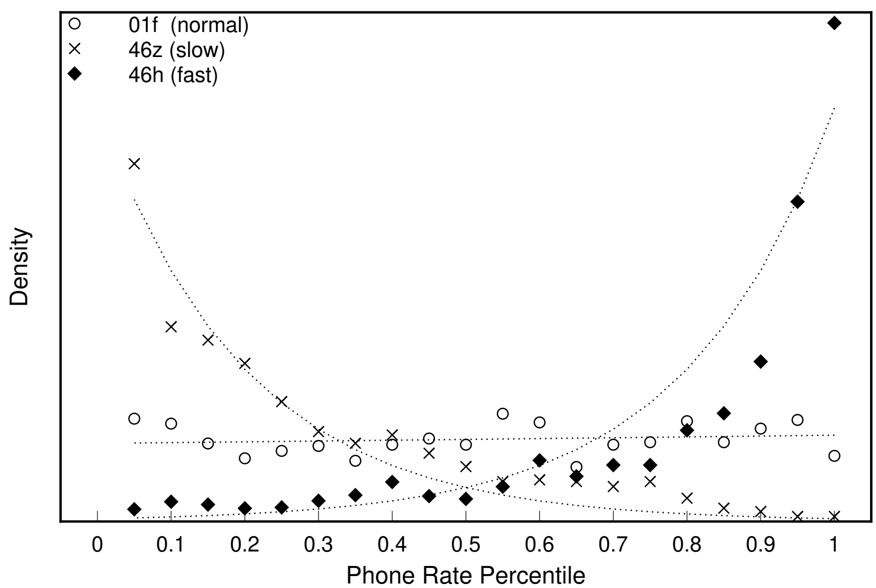

Figure 4-4 Histogram of the phone duration percentile for three different speakers. Note the exponential shapes for the fast and slow speakers. Dotted lines are the best modified exponential fit to each curve.

This distribution can be approximated reasonably well with a modified exponential PDF

$$p(x) = \begin{pmatrix} \dfrac{L}{(1-e^{-L})} e^{-Lx}, x \in [0,1] \\ 0, \text{elsewhere} \end{pmatrix} \tag{4.7}$$

where L is a parameter controlling the shape of the distribution. A family of curves for the modified exponential is shown in Figure 4-5.

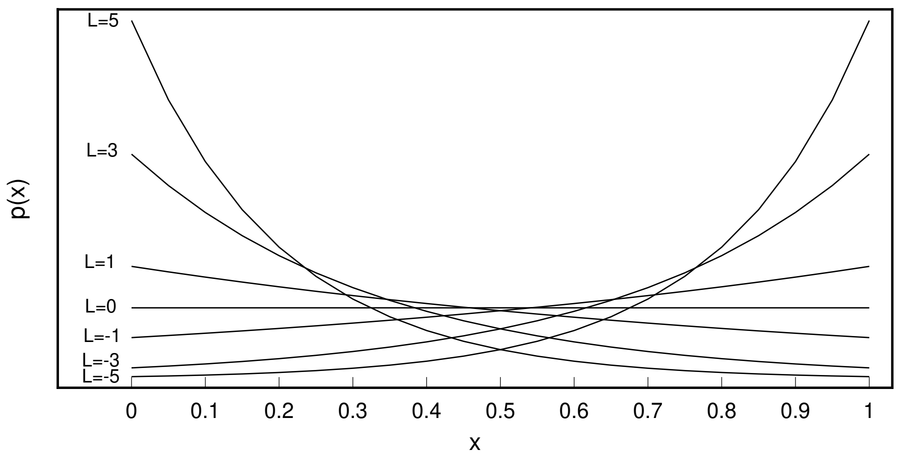

Figure 4-5 Family of curves for the modified exponential probability distribution. Each line represents a curve where the parameter L has a fixed value.

The reason for the unusual shape of this distribution is the "compressing" effect of the interval-bound percentile measure. In contrast, the unbounded z-score for a gaussian parameter is not compressed.

4.1.3.3 Sentence-Averaged Rate of Speech

Averaging the phone duration percentile of all the groups of N phones over each sentence yields a measure of the average rate of that sentence. The statistics for this sentence-wide average over the entire corpus as well as for three particular speakers are approximately Gaussian. See Figure 4-13.

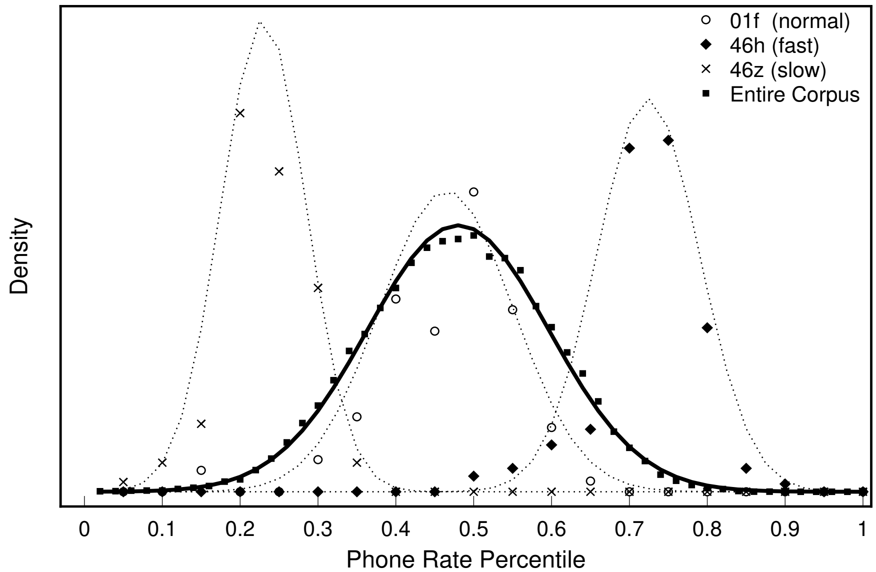

Figure 4-6 Histogram of the sentence-averaged phone duration percentile for three different speakers and the entire corpus. The histogram for approximately 100 points for each speaker are shown along with the best gaussian fit.

4.1.3.4 Word-Smoothed Rate of Speech

The use of the sentence wide average is informative about the speaker, but does not indicate sub-sentence level changes of speech rate. Instead of an arbitrary number of units, there is much evidence to support that speech rate is a **prosodic feature** of speech that is constant over the **word level** [23]. Examining the word level changes in rate percentile shows many changes in speech rate over an utterance. Of course, when the number of phones in a word is very small, the estimation of speech rate for that word becomes less reliable. All single-phone words were excluded for this reason. In addition, due to beginning and ending of sentence segmentation errors, initial and final words are excluded. The distribution of the word-smoothed phone duration percentile is shown in Figure 4-16.

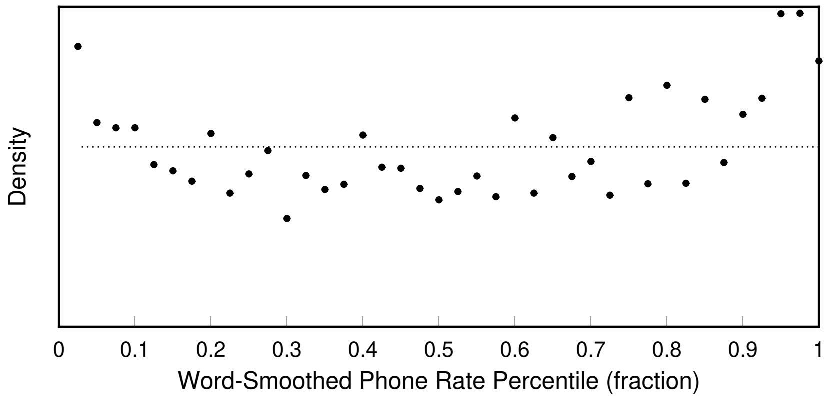

Figure 4-7 Statistics of the word-smoothed phone duration percentile. The dotted line is a uniform distribution for reference.

A graph of the word averaged rate percentile for a typical sentence in Figure 4-8 shows manifestations of known effects of speech rate. Pre-pausal lengthening is evidenced by the slowing trend before major syntactic breaks.

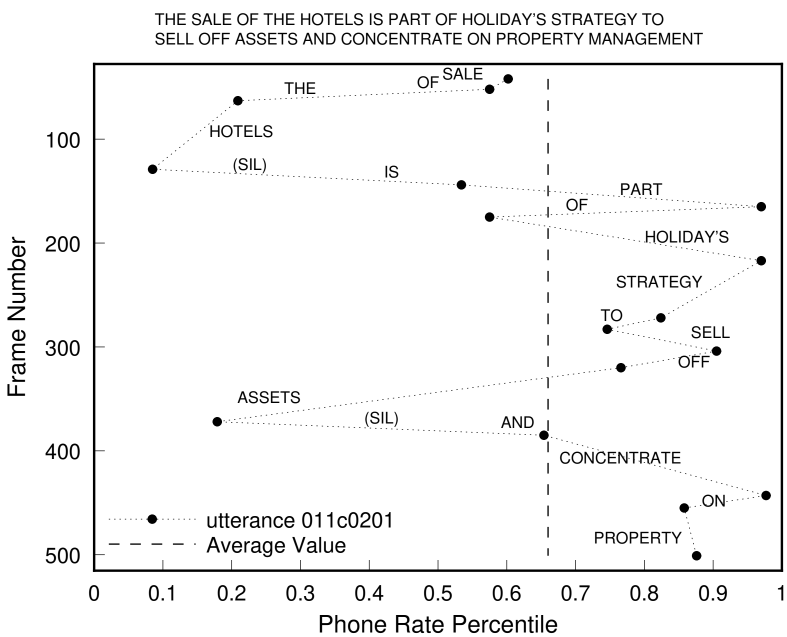

Figure 4-8 Plot of word-smoothed speech rate percentiles for utterance 011c0201. The points mark the ending frame for each word. Single phone words are omitted.

In Figure 4-9 the introduction of a new term, a proper noun, slows down speech rate. Major word-level stress and emphatic stress show a decrease in speech rate.

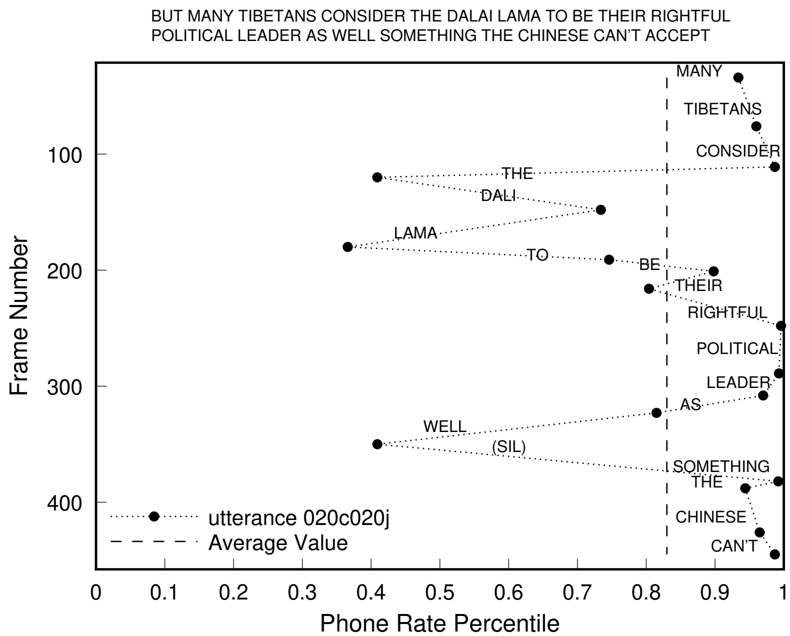

Figure 4-9 Plot of word-smoothed speech rate percentiles for utterance 020c020j. The points mark the ending frame for each word. Single phone words are omitted.

4.2 Robustness of Speech Rate Metrics

In the context of automatic speech recognition, there are two useful qualities for a speech rate metric: **text independence** and **speaker dependence**. According to O'Shaughnessy [23], "In fluent speech... speakers tend to retain a fixed speaking rate." In read speech, variations in speech rate are due more to special conditions within sentences than to global adjustments made over many sentences. A metric that exhibits text independence will report the same average speech rate for a given speaker, regardless of the text of the sentence. The metric should also be speaker dependent since the average speech rate has been shown to vary widely from speaker to speaker [7].

If sentences are grouped by speaker identity, this problem can be viewed as supervised classification. Now, the question is how well classified the data is for each particular metric. In multiple discriminant analysis [8], a set of classes is said to be optimally separated if the ratio of the inter-class scatter to the intra-class scatter is maximized.

4.2.1 Word Rate

When the word rate is broken down by speaker, as in Figure 4-13, the intra-speaker variance is often larger than the inter-speaker variance. In addition, the average intra-speaker variance is the same as the inter-speaker variance. Thus, the word rate is not robust to lexical content.

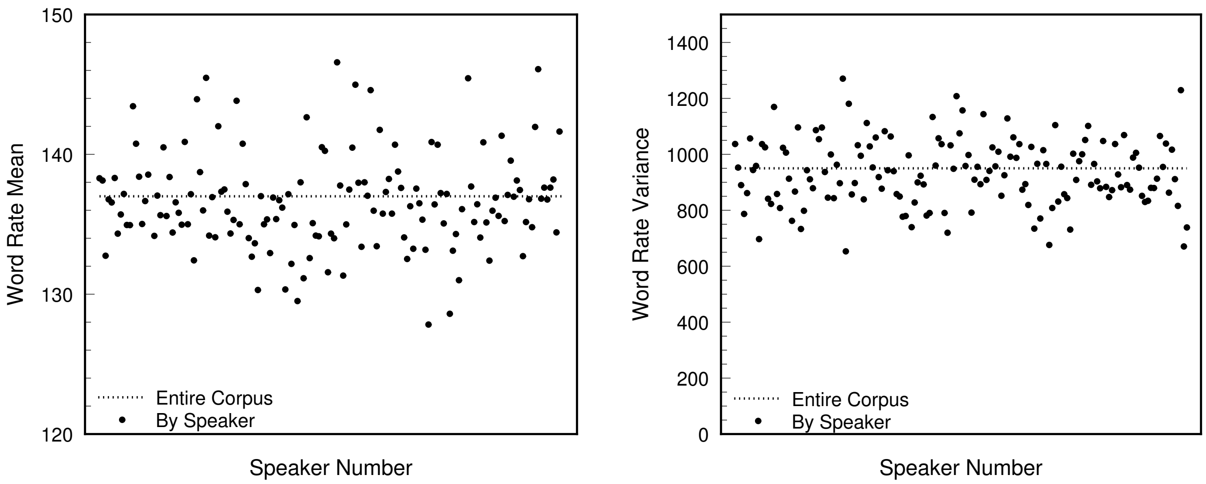

Figure 4-10 Word rate statistics of each speaker versus the entire corpus.
The ratio of inter-speaker scatter to intra-speaker scatter is 0.016.

4.2.2 Phone Rate

A breakdown of the sentence-averaged phone rate for each speaker is shown in Figure 4-11. The phone rate is more robust than the word rate since the ratio of inter-speaker scatter to intra-speaker scatter is much greater (1.3).

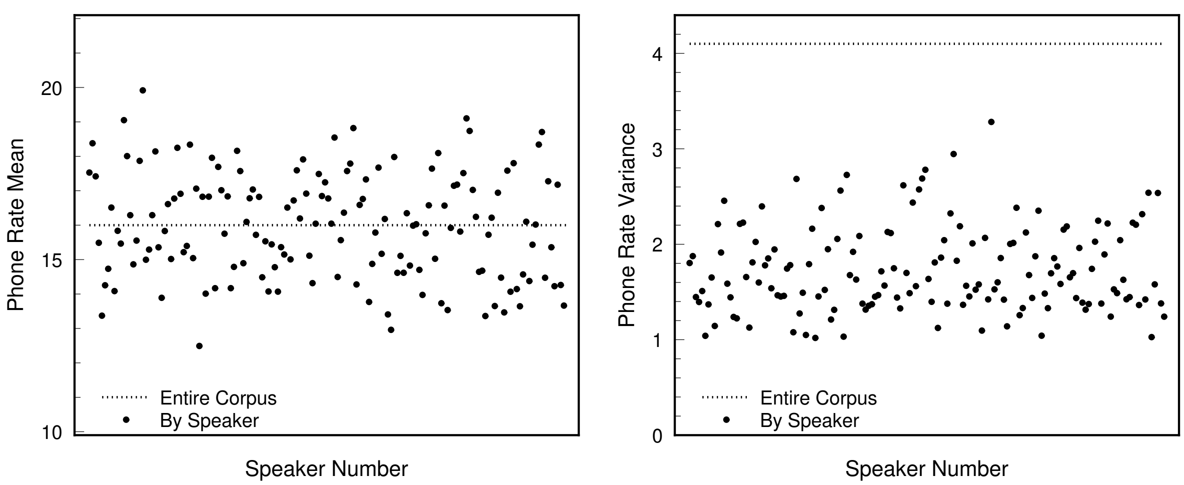

Figure 4-11 Phone rate statistics of each speaker versus the entire corpus.
The ratio of inter-speaker scatter to intra-speaker scatter is 1.3.

4.2.3 Phone Duration Percentile

Figure 4-12 is a chart of the speaker statistics for the phone duration percentile. As in the case of the phone rate, the intra-speaker variance is still large for some speakers, implying that there are local changes in the speech rate which are significant. At this point, it can be assumed that the context-dependent nature of phone duration may be giving rise to variances in the estimates of phone duration percentile. The ratio of the inter-speaker to the intra-speaker scatter is 1.7 which is greater than for phone rate.

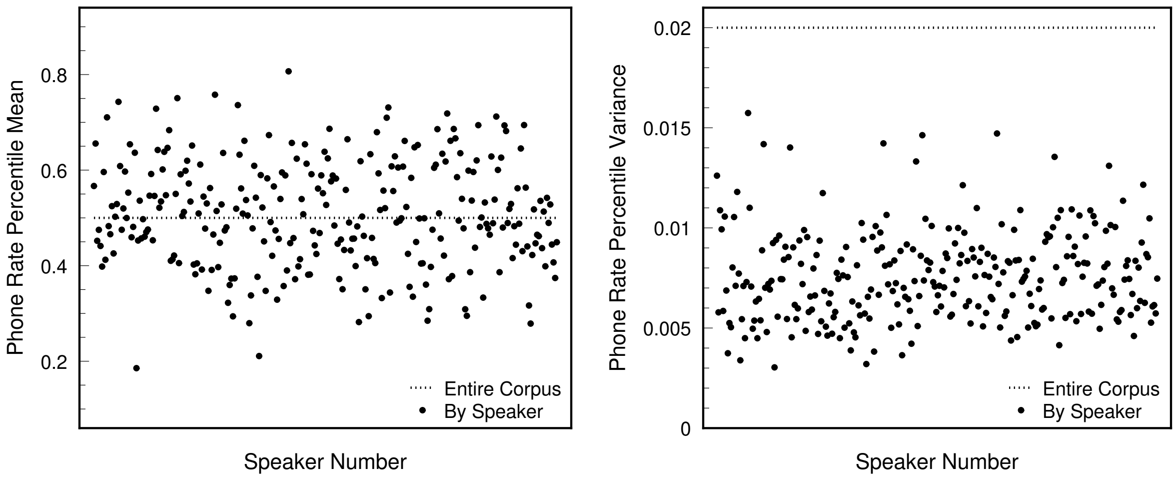

Figure 4-12 Phone duration percentile statistics of each speaker versus the entire corpus. The ratio of inter-speaker scatter to intra-speaker scatter is 1.7.

For the phone duration percentile, the 1-phone, 3-phone, and word-level smoothing windows were all applied, and statistics were taken. The rate of speech for each window was computed using the method discussed in Section 4.1.3. The distributions of sentence-average speech rate using three different smoothing windows are shown in Table 4-1. Because σ_m^2 is smaller than σ_0^2, the prior statistics for the mean and scatter of the sentence-average speech rate are indeed "sharper" than the statistics of the entire corpus.

4.3 Effects of Speech Rate on Recognition Accuracy

It is important to note what types of errors occur when the speech rate is of a particular value. For each of the three metrics discussed, recognition was performed on a large number of sentences at a variety of speech rates from the WSJ1 or WSJ01 corpora. The test sets were chosen to have no out-of-vocabulary occurrences (OOVs). The frequency of abnormal speech rates is high in the test sets to clarify their effects.

Smoothing Window	Statistics for the Entire Corpus		Statistics for All Speakers			
			Average m		Scatter s	
	Mean μ_0	Variance σ^2_0	Mean μ_m	Variance σ^2_m	Mean μ_s	Variance σ^2_s
1 phone	0.43	0.0070	**0.43**	0.0044	**0.0026**	6.5e-7
3 phones	0.50	0.020	**0.50**	0.012	**0.0076**	5.1e-6
Word	0.49	0.025	**0.51**	0.015	**0.011**	1.0e-5

Table 4-1 Statistics of sentence-average phone duration percentile in the WSJ01 Corpus.

4.3.1 Word Rate

Figure 4-13 shows recognition errors observed for subsets of the WSJ01 corpus with different word rates. It can be seen that error rate increases significantly when the word rate is greater than the mean word rate by more than two standard deviations. Utterances with a larger than average error rate due to the word rate are found only when the word rate is greater than 190 or less than 60. Approximately 3% of all utterances are in this region.

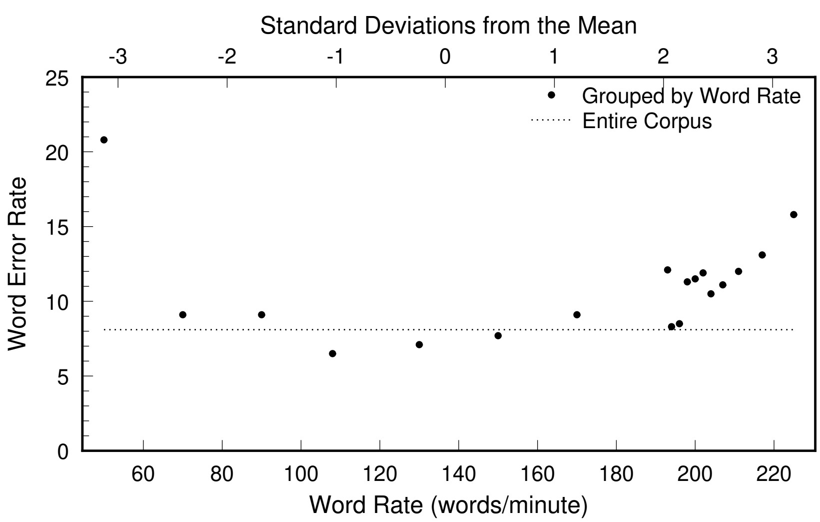

Figure 4-13 Word error rate for groups of utterances having the same word rate. Each point represents a group of 100 utterances. The dotted line is the WER for the entire corpus.

4.3.2 Phone Rate

Figure 4-13 shows recognition error rate as a function of phone rate, again based on subsets of 50 utterances of similar phone rate from the original WSJ1 corpus. In fast speech, increases in both deletion and substitution errors were responsible for the overall increase in error rate.

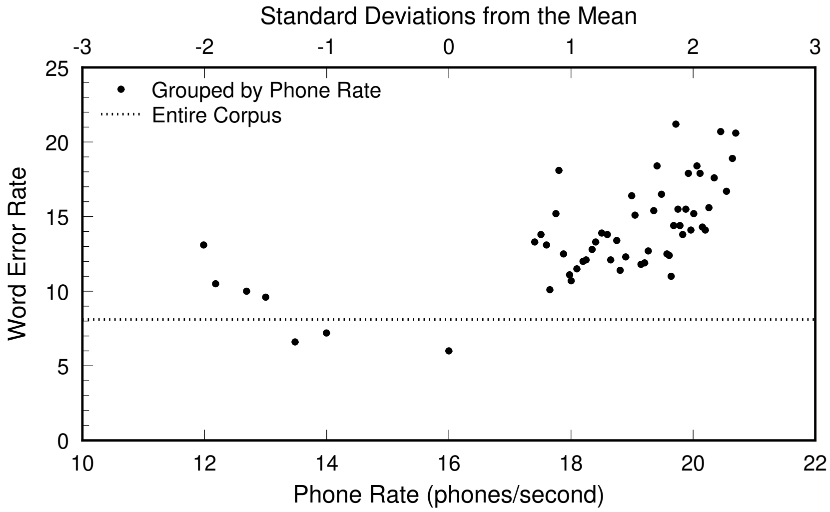

Figure 4-14 Word error rate for groups of utterances having the same phone rate. Each point represents a group of 50 utterances. The dotted line is the WER for the entire corpus.

The phone rate need only be one standard deviation above the mean for error rate to increase, while the word rate must be above two standard deviations. Utterances with a larger than average error rate due to the word rate are found when the phone rate is greater than 18. Approximately 16% of all utterances are in this region. As a result, it appears that error rate is more sensitive to phone rate than word rate.

4.3.3 Phone Duration Percentile

A test set of 1100 WSJ01 utterances from 55 speakers was selected to provide a uniform distribution of speech rate measured with the sentence-averaged phone duration percentile and the 3-phone smoothing window. When the sentence-averaged phone duration percentile is less than 1 standard deviation, the error rate is larger than average. Approximately 16% of all utterances are in this region. The response is similar to that for the phone rate metric, except that slower phone duration percentiles do not increase error rate.

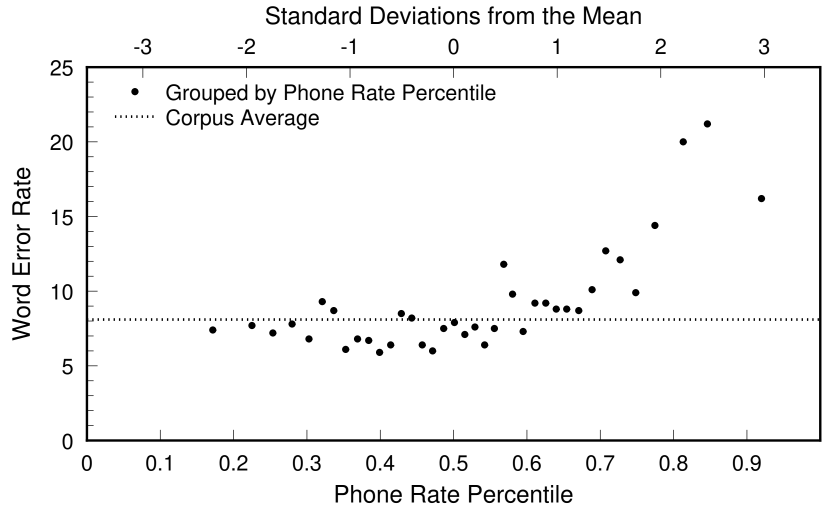

Figure 4-15 Word error rate for groups of utterances having the same sentence-averaged phone duration percentile. Each point represents a group of 50 utterances. The dotted line is the WER for the entire corpus.

4.4 Estimation of Speech Rate

It is important to be able to estimate the speech rate of a sentence without knowing the transcript since realistic compensation schemes do not have access to the text. Estimation should yield as low an error variance as possible to avoid mis-classification of speech rate.

Estimating the speech rate for any of the three metrics is very reliable. The correlation between the hypothesized and the reference word rate is 0.97 and the error variance is approximately 8. The correlation between the hypothesized and the reference phone rate is 0.93 and the error variance is approximately 1. Figure 4-16 is a scatter plot of the hypothesis and reference phone duration percentile for approximately

1000 utterances. The correlation between hypothesis score and reference score is 0.92 and the error variance is 0.004 for normal speech and 0.085 for fast speech.

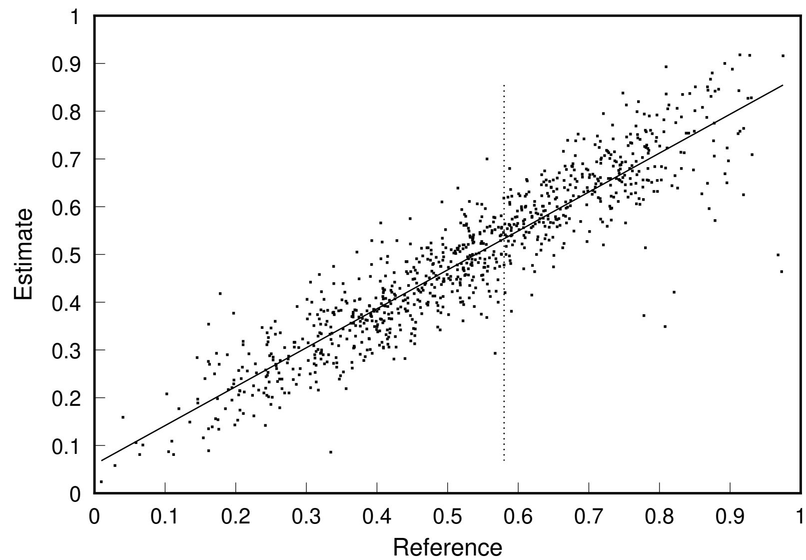

Figure 4-16 Estimated versus reference sentence-averaged phone duration percentile. Each point represents a single utterance. The solid line is the best linear fit. Overall correlation is 0.92.

4.5 Summary

Speech rate computed from CI phone durations is both robust to lexical content (text independent) and is speaker dependent. In recognition experiments, speakers among the fastest 30% of all speech had a word error rate between 50% and 150% higher than normal speech. In addition, estimates of speech rate based on baseline decoder hypothesis are accurate for this metric and could be used to blindly detect the presence of fast speech.

Chapter 5: Compensation Techniques

Degradation of recognition accuracy due to high speech rate may occur when the acoustic models and language models obtained from training with average speech fail to describe the corresponding characteristics of fast speech. In this chapter modifications are made to five components of the recognition system to compensate for the effects of high speech rate: the models of the acoustical characteristics of speech sounds, the models of the HMM state-transition probabilities, the pronunciations of words in the dictionary, the weight for combining acoustic and language evidence, and the base phone set. These procedures were selected for consideration in this work because they are relatively easy to implement without multiply retraining the speech recognition system. Since these methods unfortunately provide only limited benefit, some suggestions for more computationally demanding approaches are discussed.

In all experiments the rate of speech was measured using the phone duration percentile as defined in Section 4.3.3. In addition **fast** speech is defined as having a phone duration percentile greater than 0.6 since this is where the break in performance was demonstrated to occur. The test set in each experiment is composed of 600 fast speech utterances containing a total of 9685 words. In all experiments, the primary indication of recognition performance is the word error rate. This is defined as the total number of substituted, inserted, and deleted words divided by the total number of reference words. For a test set of this size, changes in the word error rate of 0.1% are statistically significant. Baseline performance is measured using models trained from the SI-284 WSJ01 corpus. More details about the experimental setup are described in Appendix A.

5.1 Codebook Adaptation

If the production of fast speech differs from the production of speech at a normal rate, the acoustical characteristics of the output may differ as well. In previous research, recognition accuracy has been improved through the use of VQ codebooks that were specific to gender, pitch, and environment [1][17]. To test this method Baum-Welch codebook adaptation to the fastest 1000 utterances was performed to develop rate-specific mean codebooks for fast speech.

When compared with the variance codebook vectors, the average displacement of the mean codebook vectors was found to be only 0.002 standard deviations. In comparison, adaptation to acoustic environments such as change of microphone or SNR level typically result in an average displacement of over 0.1 standard deviations.

Recognition error rate of 600 fast speech utterances using the codebooks derived from fast speech did not improve compared to the baseline error rate of 11.7%. The fact that the use of rate-specific codebooks was unsuccessful in this experiment suggests that the long-term average acoustic characteristics of normal speech and fast speech are similar. However, it is also known that codebook variations can depend on phonetic class [18], so it is possible that the use of codebook modifications based on phoneme class in addition to speech rate may be more successful.

5.2 Modification of HMM Transition Probabilities

Since faster speech is typically less carefully articulated, it is expected that recognition accuracy could be improved by modifying the representations of duration in the baseline HMMs to more closely match those of fast speech. Peterson and Lehiste (1960) showed that when speech rate increases the change in duration of vowels is greatest [27]. Forced-alignment techniques were used to confirm this observation for the WSJ01 corpus. From state and phone segmentations for all utterances, a very high incidence of extremely short (30-ms) durations in fast speech can be seen.

Figure 5-1 compares histograms of vowel durations for normal and fast speech in the WSJ01 corpus, along with the duration statistics of vowel segments in the original HMM representation for normal speech. It can be seen that the phone durations of the vowel segments of the HMMs more closely resemble normal than fast speech.

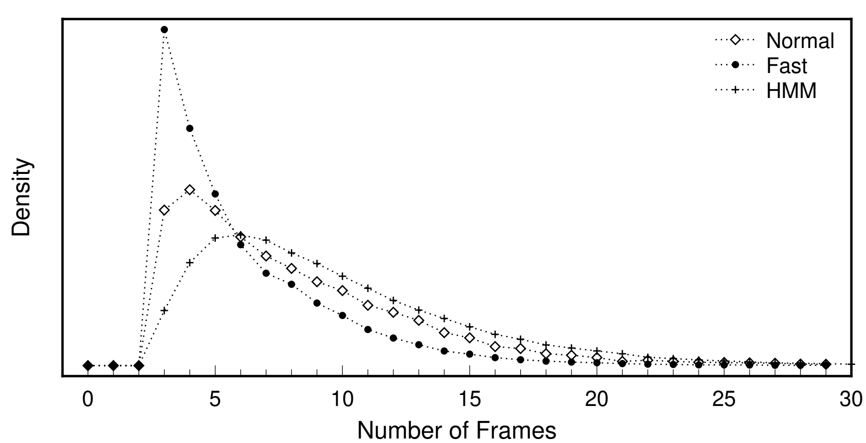

Figure 5-1 Histograms of vowel durations for normal and fast speech in the WSJ01 corpus in comparison to those derived from the HMM representation.

Although others have shown that more explicit modeling of phone duration can improve recognition accuracy [3][21][34], this would require major modifications to the recognition system trainer and to the decoder. Barring such measures, better modelling of the duration of fast speech is best accomplished by modification of the state-transition probabilities in the HMMs. In the SPHINX-II system, only monophone transition probabilities are trained because they are assumed to be relatively unimportant in recognition.

To evaluate the importance of state-transition probabilities two new sets of state-transition models were created. The state-transition probabilities in the first set were made equal to ignore any information present. The second set of models had transition probabilities adapted to the fast utterances, by counting state-transitions from segmentation of these utterances. These models were then tested on fast speech. Recognition error rates from using models with equal state-transition probabilities increased from a baseline of 11.7% to 12.2%. For models using adapted state-transition probabilities recognition the error rate dropped to 11.4%, a relative decrease of 2.6%.

Because discarding the information provided by the transition probabilities causes a reduction in performance, the probabilities are important in the recognition of speech in general. In addition, when they better match the transitions of fast speech, recognition improved. This suggests that more attention in general needs to be given towards the modelling of state-transition probabilities.

5.3 Modification of Pronunciation Dictionary

The goal of this technique is to produce lexical models for fast speech that are consistent with acoustic models trained on all speech. By their very nature, lexical models are designed to accommodate a few possible pronunciations of the same word using a basic phone set that is derived from “ordinary” speech. The linguist who produces these pronunciations desires to produce speaker-independent (SI) lexical models. This works in concert with the speaker-independent acoustic models to provide robustness to different speaker. However, as is the case with any SI models, they are less optimal than speaker-dependent (SD) models tuned to any particular speaker or group of speakers.

Most statistically based speech recognition systems such as the SPHINX-II system ignore prosodic cues such as speech rate in both training and decoding. However, the selection of pronunciations in the dictionary actually reflects the effects of rate, stress and voicing which are indeed prosodic features. As a result, systems that ignore speech rate outright are unable to model the acoustics of normal, slow, and fast speech simultaneously because these distinctions are neither labelled in training data nor accounted for in the lexical modelling.

5.3.1 Intra-Word Transformations: Rule Based Approaches

Within a word, there are many complex rules that can be applied to normal pronunciations to arrive at observed fast pronunciations. Previous studies [27][16][31][36][11][28] demonstrated that as speech rate increases, the acoustics and durations of vowels are transformed in the following manner:

- Stressed →Unstressed
- Unstressed→Schwa
- Schwa→Deleted

The rules that govern these changes depend on the phonetic context, lexical stress, accent, and many other factors. Two simple rules were applied to ascertain the viability of pronunciation transformations for fast speech. The numbers in parenthesis indicate the fraction of all pronunciations changed and the frequency of those words in the test sets:

1. Remove the schwa when it occurs before a dental consonant (1.9%, 0.76%)
2. Remove all non-initial and non-final schwas. (53.3%, 28.9%)

To model these changes for fast speech, a new set of pronunciation dictionaries was created and tested. For each rule, recognition was performed using the new dictionary, and a union of the new and original dictionary.

The first rule predictably did not improve recognition as it altered only a very small subset of words in the test set. Applied by itself, the second rule resulted in a word error rate of 22.8% which is nearly twice the baseline of 11.7%. When the second rule is combined with the baseline dictionary, the word error rate is reduced to 11.4% for a relative reduction of 2.6%.

Although they may model the pronunciation of fast speech better, the rules may have not better assisted recognition because the number of unique pronunciations decreased when they were applied. Sloboda (1995) has shown that with all other things being held constant, as the number of confusable pronunciations increases recognition errors increase as well [33]. In the baseline dictionary 10.6% of the words have the same pronunciation as at least one other word, and these words represent 23.1% of all the words occurring in the test set. In the dictionary derived from the second rule, these values are 18.5% and 35.1%. When the dictionary derived from the second rule is combined with the baseline dictionary, these values become 14.1% and 23.3%.

5.3.2 Inter-Word Transformations: Compounds

In observation of the recognition performance of fast speech, it was found that the deletion of short words is 50% higher than for other words for fast speech. Many of these deletions are of the form:

X Y -> X
X Y -> Z

where X,Y and Z are short words. For example, there were many occurrences of the following merges:

OF THE -> OF
AND A -> THE

Since it is not possible to reduce deletion errors due to merges of short words by changing their individual pronunciation, **compound-phrases** were added to the dictionary instead. Compound-phrases have slightly different pronunciations for the most frequent combinations of short words and are labeled as "OF_-THE" for example. Eide (1995) [9] has obtained a slight decrease in error rate using dictionaries with compound-phrases in spontaneous speech tasks where sequences of short words are often poorly articulated. The same cannot be said for at least one other study [5] using dictation tasks.

To evaluate this method for fast speech, a new dictionary was constructed containing 164 compound-phrases. These phrases account for 3.3% of word-pairs in the test set. When the new dictionary was used in the recognition of fast speech, many pairs of function words were labelled with their compound-phrases but the overall word error rate did not change significantly from the baseline of 11.7%.

5.4 Adjustment of Language Weight Parameter

Combining evidence from the acoustic and the language model is necessary in any speech recognition system where the two are modelled separately. In the SPHINX-II decoder the **language weight** is multiplied to the language model score. Generally, this parameter is optimized over all speakers in a test set instead of optimized for each speaker or condition.

In this experiment, the language weight was adjusted globally for the best overall performance on all fast speech, and adjusted locally for the best performance on groups of 100 fast speech utterances with similar **known** speech rate. The total test set in both cases was 600 utterances. Global optimization resulted in the word error rate dropping from the baseline of 11.7% to 10.9% for a 6.8% relative reduction. For local optimization, the word error rate dropped from the baseline of 11.7% to 10.5% for a 10.3% relative reduction.

One reason that tuning the language weight improved the performance of fast speakers may be related to how the parameter changes the number of the words permitted in the hypotheses. As the language weight decreases, less penalty is associated with adding more words to the hypotheses. For fast speakers, lower values of language weight worked best. This is sensible because in fast speech more words occur per unit of time than in normal speech.

5.5 Rate-Dependent Phone Sets

Because the short-term acoustics of fast speech have been shown to be different from the acoustics of normal speech [35][20][6][24] it is plausible that a new set of phones created specifically for modeling fast speech could improve recognition. The base phone set used in SPHINX-II was developed to provide the best representation for the most frequent pronunciations used by general American English people [13]. Since the average speech rate is a speaker dependent phenomenon, its effects are constrained to a subset of all speakers. As is the case for the pronunciation dictionary, one model cannot adequately represent both normal and fast speech simultaneously.

In this experiment, an augmented phone set was constructed as the union of the base phone set with a new phone set intended to model explicitly the acoustics of fast speech. The new fast phones were first initialized with the baseline models. Then the word-averaged phone duration percentile was used to identify the fastest 30% of all words in the corpus and the acoustics for fast phones were collected from these words. All senones were then re-clustered for this augmented phone set. In general, fewer clusters were assigned to the fast phones because of the smaller quantity of data. The total number of parameters in the system was kept constant. The training set was then run through Baum-Welch adaptation, with the acoustics from the fast words being applied to the fast phone set, and the acoustics from the normal words being applied to the normal phone set. After several iterations of Baum-Welch, the models were tested on fast speech. Because occurrences of fast words were removed from the training for normal phones and vice-versa, the phone set became **rate-dependent**.

Several pronunciation dictionaries were constructed to test the new phone sets, with entries having combinations of normal and fast pronunciations using all or some of the augmented phone set. These are summarized in Table 5-1. For each named dictionary, the phone set and the pronunciations used are listed. The

Name	Phone Set Used		Pronunciations Used		Word Error (%)
	Normal	Fast	Normal	Fast	
Baseline	N/A		N/A		11.7
Fast Only	none	all	none	all	23.4
All phones + All Pron.	all	all	all	all	13.5
All Phones + Selected Pron.	all	all	all	frequent	13.7
Selected Phones + All Pron.	all	frequent	all	all	14.3
Selected Phones + Fast Pron.	all	frequent	none	all	13.5
Selected Phones + Selected Pron.	all	frequent	all	frequent	13.7

Table 5-1 Dictionaries used to evaluate the rate dependent phone set and their performance on fast speech. **Phone Set Used** refers to which rate dependent phones where actually used in the dictionary. **Pronunciations Used** refers to which normal or fast alternative pronunciations were used. **Frequent** refers to components with an above average frequency of occurrence.

items listed as **frequent** were determined as follows:

- Unconstrained phone recognition was performed to determine what phones are most likely to be deleted in fast speech. In this experiment there was no dictionary or grammar as any phone sequence was allowable. The 20 phones (out of 50) that were deleted with above average frequency were considered as being more rate-dependent than those that were not and are defined as the frequent set.
- For the dictionaries, baseline recognition of fast speech was reanalyzed to find the words that were deleted with above average frequency (1408 of 20000). These words were considered as being more rate-dependent than those that were not. They comprise the frequent rate-dependent pronunciations.

None of the new dictionaries with rate-dependent phone sets showed an improvement over baseline recognition. In fact, many of them showed a marked increase in word error rate. The most likely source for this increase in word error rate is the decrease in the amount of training data applied to each phone. Because the number of senones was kept constant and the number of base phones was increased, there were fewer senones per phone. It is possible that increasing the size of the training set may improve recognition for this reason.

5.6 Summary

The long-term acoustics of fast speech are similar to those of normal speech. Adaptation to fast speech must be more sophisticated than simpler methods used for channel and noise compensation.

Improved temporal modeling of fast speech can be accomplished by adaptation of the transition probabilities. More sophisticated duration modeling could improve recognition of all speech.

Fast speakers use different pronunciations for many words. Lexical modeling that is specific to fast and normal speech can improve recognition.

The proper combination of acoustic and linguistic models is critical for obtaining optimal performance in fast speech. As a result, fine tuning of the language weight can improve recognition.

Since the short-term acoustics of fast and normal speech are different, some implementation of rate-dependent modeling at the senonic level should improve recognition.

Chapter 6: Conclusions

Very simple measures such as word or phone rate are not suitable for detection of both long-term and short-term speech rate since they are sensitive to the lexical content of the speech. In contrast, the phone duration percentile, a comparison of measured versus expected phone duration, is robust to lexical content and consistent with previous findings about the statistics of long-term and short-term speech rate. Estimation of the speech rate using this metric from decoder hypotheses is reliable.

In measurements of recognition performance of speech with similar rate, it is confirmed that speech rate affects the performance of automatic recognition. Increases in word error rate from 50% to 150% are typical for the fastest 30% of read speech. In spontaneous speech the increases may be even higher.

Compensation for the effects of fast speech using simple adaptation methods in various components of the recognition system proves to be elusive. None of the methods described in this report provided significant improvements, but some important conclusions can be drawn by their failure:

- Long term acoustics of fast speech are the same as normal speech. As a result, methods which adapt to speakers using simple adaptation of the means of the features will not work.
- Temporal models of the events of normal speech are inconsistent with models of fast speech. There is much research that supports the use of duration modeling that is specific to normal and fast speech.
- Lexical models of normal speech are not the same as models for fast speech. Special dictionaries which account for the differences in pronunciation of words in fast speech may recover this.
- The best balance between acoustic and linguistic probability is sensitive to speech rate. As speech rate changes, optimizing the weight given to each probability results in improved performance. This technique may work for other qualities besides speech rate as well.
- There are some unrecoverable acoustic losses when speech rate is fast. Very detailed modeling at the phonetic level cannot be used to improve recognition. This confirms findings by many others about the qualities of fast speech.

As the number of applications of speech recognition systems in real environments increases, so will the expectations for the robustness of these systems to unusual speakers. Speech rate represents only one of many characteristics of such speakers which cause degradation.

Appendix A: The WSJ01 Corpus

A detailed description of the goals and elements of the Wall Street Journal corpus can be found in [26]. For the experiments in this work, both parts 0 and 1 were used in the training and testing components and together are named the WSJ01 corpus. The motivation of the WSJ01 corpus is to develop speaker independent recognition of continuous speech for dictation and other types of fluent speech tasks.

The speaker independent component of the corpus is composed of 284 speakers of American English, each speaking between 50 and 100 randomly chosen complete sentences from Wall Street Journal articles into a studio quality microphone. The recording was supervised so that only fluent speech was accepted. There are a total of approximately 35000 utterances in the corpus of varying lengths from 1 to 25 seconds.

Although the entire corpus contains only approximately 15000 unique words, the lexicon used in recognition tests contains approximately 20000 unique words derived from statistical analysis of Wall Street Journal articles preceding the recordings. This means that there are some words occurring in training sentences which are not in the testing lexicon. For the experiments in this work, utterances containing such out-of-vocabulary (OOV) words were excluded.

In this report, training material was used for testing. In a very simple task this would not reflect the true performance of the system. However, in a task with a large number of speakers and a large vocabulary, the effects of testing on training material are negligible. Expected performance with the SPHINX-II recognition system varies anywhere from 7-12% depending on the size of the lexicon and language model used.

References

[1] A. Acero [1993], *Acoustical and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, MA.

[2] A. Anastasakos, R. Schwartz, H. Shu [1995], "Duration Modeling in Large Vocabulary Speech Reconition," ICASSP, Vol. 1, pp. 628-631.

[3] D. Burshtein [1995], "Robust Parametric Modeling of Durations in Hidden Markov Models," ICASSP, Vol. 1, pp. 548-551.

[4] W. Campbell [1988], "Extracting Speech-Rate Values From a Real-Speech Database," ICASSP, Vol. 1, pp. 683-686.

[5] L. Chase, R. Rosenfeld, A. Hauptmann, M. Ravishankar, E. Thayer, P. Placeway, R. Weide, C. Lu [1995], “Improvements in Language, Lexical, and Phonetic Modeling in SPHINX-II,” Proceedings of the Spoken Language Systems Technology Workshop.

[6] W. Cooper, C. Soares, A. Ham, K. Damon [1983], "The Influence of Inter- and Intra-Speaker Tempo on Fundamental Frequency and Palatalization," JASA, Vol. 73, pp. 1732-1730.

[7] T. Crystal, A. House [1982], "Segmental Durations in Connected Speech Signals: Preliminary Results," JASA, Vol. 72, pp. 705-716.

[8] R. Duda, P. Hart [1973], *Pattern Classification and Scene Analysis*, J. Wiley & Sons, New York, NY.

[9] E. Eide, Personal communication [1994].

[10] E. Eide, H. Gish, P. Jeanrenaud, A. Mielke [1995], "Understanding and Improving Speech Recognition Performance Through the Use of Diagnostic Tools," ICASSP, Vol. 1, pp. . 221-224.

[11] T. Gay [1978], "Effect of Speaking Rate on Vowel Format Movements," JASA, Vol. 63, pp. 223-230.

[12] J. Godfrey, E. Holliman, J. McDaniel [1992], "SWITCHBOARD: Telephone Speech Corpus for Research and Development," ICASSP.

[13] X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, R. Rosenfeld [1993], "The SPHINX-II Speech Recognition System: An Overview," Computer Speech and Language, Vol. 7, pp. 137-48.

[14] X. Huang, Y. Ariki, M. Jack [1990], *Hidden Markov Models for Speech Recognition,* Edinburgh University Press, Edinburgh, UK.

[15] M.-H. Hwang [1993], “Subphonetic Acoustic Modeling for Speaker Independent Continuous Speech Recognition,” Ph.D. Thesis, School of Computer Science, Carnegie Mellon University.

[16] I. Lehiste [1970], Suprasegmentals, Cambridge, Mass., M.I.T. Press.

[17] F.-H. Liu [1994], “Environmental Adaptation for Robust Speech Recognition,” Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University.

[18] F.-H. Liu, P. Moreno, R. Stern, A. Acero [1994], “Signal Processing for Robust Speech Recognition,” Proceedings of the Spoken Language Systems Technology Workshop.

[19] P. Luce, J. Charles-Luce [1985], "Contextual Effects on Vowel Duration, Closure Duration, and the Consonant/Vowel Ratio in Speech Production," JASA, Vol. 78, pp. 1949-1957.

[20] J. Miller, T. Baer [1983], "Some Effects of Speaking Rate on the Production of /b/ and /w/," JASA, Vol. 73, pp. 1751-1755.

[21] M. Monkowski, M. Picheny, P. Rao [1995], "Context Dependent Phonetic Duration Models for Decoding Conversational Speech," ICASSP, Vol. 1, pp. 528-531.

[22] P. Moreno, D. Roe, P. Ramesh [1990], "Rejection Techniques In Continuous Speech Recognition Using Hidden Markov Models," Signal Processing V: Theories and Applications, pp. 1383-1386.

[23] D. O'Shaughnessy [1995], "Timing Patterns in Fluent and Disfluent Spontaneous Speech," ICASSP, Vol. 1, pp. 600-603.

[24] D. Ostry, K. Munhall [1985], "Control of Rate and Duration of Speech Movements," JASA, Vol. 77, pp. 640-648.

[25] D. Pallett, et. al. [1994], "Be Sure to Read the Fine Print: II," Proceedings of the Sopken Language Technology Workshop.

[26] D. Paul, J. Baker [1992], "The Design of the Wall Street Journal-based CSR Corpus," Proc. DARPA Speech and Natural Language Workshop.

[27] G. Peterson and I. Lehiste [1960], "Duration of Syllable Nuclei in English," JASA, Vol. 32, pp. 693-703.

[28] R. Port [1981], "Linguistic Timing Factos in Combination," JASA, Vol. 69, pp. 262-273.

[29] L. Rabiner, B.-H. Juang [1993], *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, NJ.

[30] R. Rosenfeld [1989], “Adaptive Statistical Language Modeling: A Maximum Entropy Approach,“ Ph.D. Thesis, School of Computer Science, Carnegie Mellon University.

[31] L. Shockey [1973], Phonetic and Phonological Properties of Connected Speech, Ph.D. Thesis, Phonetics and Phonology, Ohio State University.

[32] M. Siegler, R. Stern[1995], “On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems,” ICASSP, Vol.1.

[33] T. Sloboda [1995], "Dictionary Learning: Performance Through Consistency," ICASSP, Vol. 1, pp. 453-455.

[34] N. Suaudeau, R. Andre-Obrecht [1994], "An Efficient Combination of Acoustic and Supra-Segmental Informations in a Speech Recognition System," ICASSP, Vol. 1, pp. 65-68.

[35] B. Tuller, K. Harris, J. Kelso [1982], "Stress and Rate: Differential Transformations of Articulation," JASA, Vol. 71, pp. 1534-1543.

[36] N. Umeda [1975], "Vowel Duration in American English," JASA, Vol. 58, pp. 434-445.

[37] N. Umeda [1977], "Consonant Duration in American English," JASA, Vol. 61, pp. 846-858.

[38] G. Weismer, A. Fennell [1985], "Constancy of (Acoustic) Relative Timing Measures in Phrase-Level Utterances," JASA, Vol. 78, pp. 49-57.