Contents

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Making Automatic Speech Recognition More Robust to Fast Speech

Nikki Mirghafori, Eric Fosler, and Nelson Morgan

TR-95-067

December 1995

Abstract

Psychoacoustic studies show that human listeners are sensitive to speaking rate variations [32]. Automatic speech recognition (ASR) systems are even more affected by the changes in rate, as double to quadruple word recognition error rates of average speakers have been observed for fast speakers on many ASR systems [24]. In this work, we have studied the causes of higher error and concluded that both the *acoustic-phonetic* and the *phonological* differences are sources of higher word error rates. We have also studied various measures for quantifying rate of speech (ROS), and used simple methods for estimating the speaking rate of a novel utterance using ASR technology. We have implemented mechanisms that make our ASR system more robust to fast speech. Using our ROS estimator to identify fast sentences in the test set, our rate-dependent system has 24.5% fewer errors on the fastest sentences and 6.2% fewer errors on all sentences of the WSJ93 evaluation set relative to the baseline HMM/MLP system. These results were achieved using some gross approximations: adjustment for one rate over an entire utterance, hand-tweaked rather than optimal transition parameters, and quantization of rate effects to two levels (fast and not fast).

1 Motivation

Anyone who has attended a public auction knows that there are differences in speaking rate between speakers. Miller et al. [16] have shown that the articulation rate varies quite considerably *within* and across speakers. These rate alterations modify the acoustic fine structure of individual syllables and affect properties that convey segmental information for both consonants and vowels [27]. Furthermore, listeners are extremely sensitive to these variations and they treat the segmentally relevant acoustic properties in a rate-dependent manner [32]. Fast speaking rates can make speech hard to understand for people (especially for the elderly). There have even been attempts to make engineering products that slow down fast speech [29].

ASR systems, even more than people, are sensitive to rate of speech (ROS) differences. For example, in a recent National Institute of Standards and Technology's (NIST) evaluation of the Wall Street Journal (WSJ) task in November 1993, all the participating systems had about 2-3 time higher word error rates on the two fastest speakers than on the normal speakers (see figures in [24] and Figure 1).

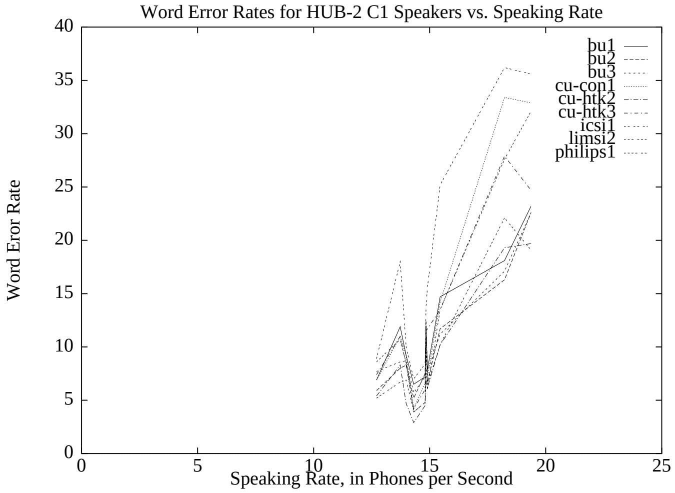


Figure 1: Rate of speech vs. word error rate for WSJ0-93 5K evaluation set. Each point represents one of the ten test speakers.

In an earlier NIST evaluation of the Resource Management task (RM) in September of 1992, this strong ROS effect was also observed. The participating systems in that evaluation had 2-4 times more error on the fastest (and one of the slowest[1]) speakers [25]. This observation naturally raises the following two questions: why do ASR systems perform significantly worse on fast speakers? And what can we do to alleviate these problems? Although fast speech has caused problems for ASR systems for some time, this issue has not been received much attention in the ASR literature. This work attempts to provide some preliminary answers to these questions.

[1] Although very slow speakers can also have high error rates, in this work we have limited our investigation to fast speakers.

In Section 2, we discuss work by others in this area. In Section 3, we briefly describe ICSI's basic ASR technology and the databases we use for this study. In Section 4, we discuss different criteria for calculating ROS, and in Section 5 we report on the experiments we have performed to decide which method most consistently characterizes ROS. In Section 6 we report our analysis of fast speech, and in Section 7 we discuss the mechanisms we have implemented for making our ASR system more robust to fast speech. In Section 9 we provide a brief summary of our work, and conclude the paper by suggesting future directions in Section 10.

2 Related Work

Although speaking rate has been a problem for ASR systems for some years, there has been very little attention devoted to this topic. The only other previously published work on this issue that we know of is that of Siegler and Stern [31]. In their work, they proposed measuring ROS based on phones/second using a mean of rates formula[2]. They used forced alignment on the correct word transcription to determine the phone segmentation and durations, and implemented three methods for compensating for fast speech errors in the WSJ1 corpus:

1. *Modification of the acoustic models*. Siegler and Stern developed a rate-specific codebook by performing Baum-Welch codebook re-estimation for fast speech. The performance with this adapted codebook did not improve for fast speech compared to the baseline.

2. *Modification of HMM state transition probabilities*. They observed that vowel duration, for example, is different for normal and fast speech. The transition probabilities of the word models were adapted to the fastest 1000 sentences in WSJ1; improvement for the fast test sentences was about 4-6%.

3. *Modification of the pronunciation dictionary: intra- and inter-word transformations*. Since the deletion of unstressed vowels are common in fast speech, the recognition dictionary was changed to eliminate the *schwa* between two consonants, as well as all the non-initial and non-final *schwas*. Neither of these changes significantly changed the overall word error rate. They also observed that function words such as THE, AND, TO, A, OF, IN, THAT, WERE, ARE and I represented 55% of all word deletions errors, even though they only represented 20% of the words in the transcripts. Since 33% of the merges were of the form “X Y” -> “X” and “X Y” -> “Z”, they added compound words such as “IN_THE” to the dictionary with a different pronunciation than each word separately. The recognition accuracy did not improve with this modification, either.

[2] This method is discussed in Section 4.4.

To estimate the ROS of a novel utterance without knowing the transcript *a priori*, they proposed using the recognition system's hypothesis containing alignment information. Their estimate of the ROS was monotonically related to the ROS calculated with the correct sentence transcription, although it was negatively biased. All their reported improvements were with the ROS calculated given the correct sentence transcriptions, however.

The ideas discussed in this report were developed independently in the same time frame as the work of Siegler and Stern [31].

3 System and Databases

For our study, we use the HMM/MLP Hybrid System and the TIMIT and the Wall Street Journal (WSJ) databases.

3.1 HMM/MLP Hybrid System

We use ICSI's hybrid HMM/MLP speech recognition system (explained in [2]). The main idea of the HMM/MLP method is to train a multi-layer perceptron (MLP) (typically using a relative entropy error criterion) for phonemic classification; such a net can be used as an estimator of posterior class probabilities and, when divided by class priors, can estimate scaled likelihoods. For recognition, we use a decoder called Y0 (described in [28]), which uses a single density per phone HMM with repeated states for a simple durational model. ICSI used this system to participate in the WSJ 93 and RM 92 NIST evaluations, and the behavior of the system on the fastest speakers was similar to that of the other systems (Figure 1).

Since similar rate of speech effects have been observed for recognizers incorporating mixtures of Gaussians [24, 25, 31], we think it likely that the conclusions of our work will be useful in those systems as well.

3.2 The TIMIT Database

The TIMIT [4, 13, 21] read-speech database was collected by Texas Instruments (TI) and was automatically phonetically labeled, and later hand-checked by the students at the Massachusetts Institute of Technology (MIT). It is available on CD-ROM from NIST, and comprises two subsets: the TRAIN and the TEST set. There are a total of 630 speakers (462 TRAIN, and 168 TEST), each uttering 10 sentences. There are three sentence types: SA sentences (2 per speaker), SX (5 per speaker) and SI (3 per speaker). The two SA sentences are the same for all speakers and are often used for calibration but not for training; the SX sentences are specifically designed to provide a good coverage of pairs of phones, and the SI sentences are selected from existing written sources to add diversity to the corpus. Each of the SX sentences is spoken by seven speakers but each SI sentence occurs only once. There are almost twice as many male as females in the TIMIT database.

TIMIT is a particularly good test-bed for our experiments because it is hand-labeled. If a database is not hand-labeled (in case of WSJ, for instance) we must generate phonetic labels through a forced alignment procedure. However, this may introduce a bias into the experiments: if the "normal" word (or phone) models do not match fast speech (due to phone omission, for example), the phonetic labels will not be 100% accurate. With hand-labeled TIMIT, this is not as much of a concern. Even though hand-labels may differ from one expert to another, they still provide the best approximation to the "ground truth".

Finally we must note that TIMIT is generally considered a phone recognition task as opposed to a word recognition task. Not only is the training set too small to reliably estimate a grammar for the test set, but there are also many out of vocabulary words in the test set that do not appear in the training set.

3.3 The Wall Street Journal/North American Business News Database

This database was recorded by various sites, and is available on CD-ROM from NIST [26]. In our study, we use WSJ0, which is a subset of the training data. There are 84 training speakers in the WSJ0 short-term training set (a.k.a. SI-84), each uttering 50 or 100 sentences. The average length of each sentence is about 7.4 seconds, and there are 15.3 hours of training data in SI-84. The training set is relatively gender-balanced, and each speaker reads a different set of sentences from the Wall Street Journal daily newspaper.

An advantage of the WSJ database for our experiments is that it is a continuous word recognition task, as opposed to TIMIT which is generally regarded as a phone recognition task. Also, the size and the phonetic variability of WSJ corpus allows us to experiment with duration as well as pronunciation modeling for fast speech.

4 Issues in Measuring Rate of Speech (ROS)

To improve robustness to speaking rate, we first need a consistent measure for quantifying speaking rate. In the course of our study, we noticed a lack of consensus in the literature on how to quantify speaking rate. It has been our experience that choosing one ROS metric over others can lead to significant differences in experimental results.

Various measures of ROS have been used by different researchers. Crystal and House [3] use the total reading time of the text to distinguish fast and slow talkers. Similarly, Ohno and Fujisaki [23] calculate a local speech rate with respect to a given target utterance. These measures are only useful if the talkers are reading identical texts, and are useless outside the laboratory environment. Clearly, one important requirement of our ROS measure is the ability to calculate it for a novel utterance.

In the next few subsections, we will discuss the following issues for calculating ROS:

- Treatment of mid-sentence silences
- Granularity of calculating ROS

- Units of ROS
- Formula for calculating ROS
- Using ASR technology to estimate ROS

In the next section, we will report on our experiments to determine the most reliable method of ROS estimation.

4.1 Treatment of Mid-Sentence Silences

Speaking rate is composed of two elements: the rate at which the speech itself is produced, or *articulation rate,* and the number and duration of pauses in the utterance, known as *pause rate* [16]. The psychoacoustic studies have established that for human listeners, the perceived changes in speaking rate that occur both within and across speakers are largely due to changes in pause rate, with the articulation rate varying relatively less [5, 6]. The articulation rate nevertheless varies quite considerably within and across speakers [16].

Although both pause rate and articulation rate are important components of speech rate, we are concerned that the duration of the silence periods may be dependent on factors other than speech rate. We argue for defining the rate of speech (ROS) as the articulation rate alone. To justify this, in Section 5 we will compare the reliability of a measure which preserves mid-sentence silences with one that excludes them.

4.2 The Granularity of Calculating ROS

Should ROS be calculated per speaker or per sentence? The advantage of the former is that it allows the grouping of speakers into “fast” and “slow” speakers, which is intuitive. That is, it is consistent with the human notion of categorizing speakers as fast (e.g., auctioneers) and slow (e.g., one’s grandmother). The disadvantage is that for a given speaker, the ROS varies considerably *across* and *within* sentences [16]. Imagine a speaker who, at the beginning of the recording session, succeeds in sustaining a normal speed, but by the end of the session, speeds up his/her speaking rate due to impatience or fatigue. Labeling all the sentences of this speaker as “medium fast” may cause anomalies in the observations. This problem is more of an issue for a corpus such as WSJ0, where each speaker utters about 50-100 sentences, than for TIMIT, where each speaker utters 10 sentences. In Section 5.2.1 we will measure the intra-speaker ROS variabilities for TIMIT.

Miller et. al. [16] have shown that there are ROS variations even within an utterance. There is some evidence that commonly used words, or *function words* are pronounced the most carelessly [12]. Having listened to many sentences, we have also noticed many rate changes in the middle of the sentence, particularly for the common expressions. One may argue that perhaps a sentence is too coarse of a unit for the ROS calculation, and ROS should be calculated either per phone, per syllable, or per 1 second segments. One problem with per phone measures is the following: the average duration of phones varies greatly, for example for TIMIT the average duration of /ow/ is 128 msec while the average duration of /k/ is 50 msec. Calculating an instantaneous rate for each of these phones may be misleading. However, since the phone rate averaged over an entire utterance provides a smoothed, though coarse, measure, we decided this was a good choice for a first attempt. In any case, for applications which require phone rate determination, we suggest a way to calculate this rate more reliably in Section 10. For this work, we have chosen a sentence level granularity, because not only is it a well defined unit providing a good approximation to the overall speed of the sentence, but it may well be sufficient for many ASR applications, particularly those with short utterances.

4.3 Units of ROS

In some studies [24, 34] the ROS has been measured based on words/minute (or per second). Although words/second is a simpler unit to calculate, it is coarser than phones/second and may cause inaccuracies. Consider the two perennial favorite examples of speech researchers: “How to wreck a nice beach” and “How to recognize speech”. If we use words/second as unit, these two sentences which have nearly identical phonetic structure, spoken at the same speaking rate, will be labeled with widely varying ROSs.

Choosing words/second as ROS units is particularly problematic if ROS is measured on a per sentence basis. Siegler and Stern [31] show a correlation of 0.50 between ROS based on words/second and phones/second for WSJ1 training sentences. The correlation between words/second and phones/second measure increases when more than one sentence is used for ROS calculation; in other words, when we calculate an average over a larger number of words, according to the law of large numbers, we get an estimate that is closer to the real mean. We have observed a correlation of 0.75 between these two metrics when ROS is measured for eight sentences of TIMIT (per speaker). The correlation between these two metrics improves to 0.99 when we used 40 sentences (per speaker) of WSJ0 for ROS measurement.

In many psycho-acoustic experiments, syllables/second has been used as the ROS unit [5, 15, 7]. Since automatic labeling in ASR systems is often based on phones and not syllables, and since phones are of even finer granularity than syllables, phones seem to be the most logical choice for ROS calculation unit for common ASR systems at this time.

4.4 Formula for Calculating ROS

There are (at least) two ways to calculate the rate of speech of an utterance. One measure is the inverse of mean duration (IMD), where the total number of phones is divided by the total duration of the utterance [19] as in:

$$ROS_{IMD} = \frac{n}{\sum_i duration_i} \tag{1}$$

where n is the total number of phones, and $duration_i$ is the duration of each phone i in the sentence. The second measure is the mean of rates (MR) formulation, where first an ROS for each phone in the sentence is calculated, and then these phone rates are averaged to get the ROS for the utterance [31], that is:

$$ROS_{MR} = \frac{\sum_i rate_i}{n} \tag{2}$$

where $rate_i$ is defined as $\frac{1}{duration_i}$ for each phone.

As we noted in Section 4.2, the average duration of phones vary widely. The MR measure accentuates the differences between the average phone duration and, in case of very short phones, drastically boosts the ROS. In other words, the MR measure is dominated by the high instantaneous rate of short phones, while the IMD measure is relatively unaffected. We discuss the merits of these two methods in Section 5.

4.5 Using ASR Technology to Estimate the ROS

In Section 4.3 we discussed the merits of calculating ROS based on words/second unit vs. phones/second unit. Since we have chosen phones/second as units, we need to know the number of phones and their duration for each sentence. Unlike TIMIT, most ASR corpora are not phonetically hand-labeled, so we need another method to determine the phonetic labeling. Forced alignment is the method most commonly used for this purpose. Given the correct word level transcription of the sentence, we can use the forced alignment method to estimate the number and the duration of phones in the sentence. If we have multiple pronunciations for a particular word, or if there is a mismatch between the word model and actually what was said (due to phone omission, for example) the phonetic duration estimation may not be accurate. Since we are calculating ROS over the whole utterance, minor inaccuracies will not have a strong effect.

How can we estimate the ROS of sentences for which we do not have the correct word level transcription? There are (at least) two possible options. One is to perform *word* recognition on the novel utterance and use the hypothesized word transcription for forced alignment (also suggested by [31]). The advantage of this method is that we can rely on higher level knowledge (i.e., language model) to get a more accurate phonetic segmentation. One drawback may be that we enforce a particular pronunciation of a word, even if the "fast" pronunciation is different from the normal word-model due to phone omission, for example. Another drawback is that incorrect word recognition can lead to the incorrect phonetic segmentation. The second option is to perform *phone* recognition for the novel utterance and use the state transition information (or if a particular decoder does not provide state transition information, forced alignment on the phone string hypothesis may be used). The advantage of this method is that we can estimate the ROS for any novel utterance, even if we do not have a word model to represent it. Another advantage is that substitution errors in the phone classification do not affect the ROS measure. The drawback of both of the above methods is that their accuracy depends on the accuracy of the ASR system, which may be poorer for rapid speech. We will report our study of these methods in Section 5.

5 Experiments in Choosing an ROS Measure

5.1 Measurements

As we have discussed so far, there are many criteria for choosing an ROS measure, and we need to determine the measure with the most consistency and reliability. Since TIMIT is the only corpus that we have available that is both phonetically hand transcribed and phonetically rich[3], we use it for determining the ROS measure of choice. In the next subsections we discuss the method we used to measure ROS based on discussions in Section 4.5.

5.1.1 Calculating ROS from Phonetic Hand Transcription

For all 5040 TIMIT training and testing sentences, we used the phone transcription (*.phn) files to calculate the ROS. These files mark the beginning and ending sample number of each phone, as well as the phonetic assignment. Note that TIMIT has been sampled at 16KHz, so $num(secs) = \frac{num(samples)}{16000}$. For example, the phonetic hand transcription for sentence mtcs08-si1972, with the word transcription "Perfect he thought," is:

```
0 2180 h#
2180 3120 p
3120 4678 er
4678 6070 f
6070 7160 ix
7160 8710 kcl
8710 9360 t
9360 10360 pau
10360 11540 hh
11540 12440 iy
12440 14230 th
14230 17080 ao
17080 21420 tcl
21420 25760 h#
```

Table 18 in the Appendix explains the phonetic transcription symbols. Note that `h#` is the phonetic label for the beginning and ending silence and `pau` is the phonetic label for middle silence or pause. Since the beginning and ending silence duration contain no inherent information about the speech, we always exclude them for ROS calculation. Based on this convention, the above sentence has 12 total phones, and 11 non-mid-silence phones. We calculated the ROS for the entire TIMIT train and test sets with and without the mid-silences, using both the IMD and MR formulas discussed in Section 4.4. These values for our sample sentence are shown in Table 1.

[3] By phonetically rich we mean that there are many instances of a phone appearing in different phonetic contexts, as opposed to, say, a digits task, where phone /m/ does not appear at all.

The ROS Info for Phonetic Hand Transcription Method		
	With Mid-silences	Without Mid-silences
Num. Phones	12	11
Duration (secs)	1.20	1.14
ROS_{IMD}	9.98	9.65
ROS_{MR}	12.83	12.54

Table 1: The ROS for sentence mtcs08-si1972 from the TIMIT training set, using the phonetic hand segmentation.

5.1.2 Calculating ROS from Correct Word Transcription

In order to determine the phonetic segmentation using word transcriptions, we used a forced Viterbi alignment. The forced Viterbi alignment is a dynamic programming algorithm that calculates a phonetic segmentation of the utterance given a particular word transcription. The phonetic likelihoods needed for the alignment procedure were generated by running a feed-forward pass for the TIMIT extracted features (PLP12 [8] and energy feature as well as their deltas) through a multi-layer perceptron (MLP), which was previously trained and cross-validated on similar features extracted from the TIMIT hand segmented data. The MLP had 1000 hidden units, 61 outputs (one for each phone), and 234 inputs (26 inputs * 9 frame window); it was trained using a relative entropy error criterion to estimate the phonetic posterior probabilities. The phonetic posterior probabilities were divided by the phonetic priors to obtain the phonetic likelihoods that were used in the forced Viterbi alignment.

The forced Viterbi alignment finds a phonetic segmentation for each utterance. For our example sentence, mtcs08-si1972, this alignment is:

```
10 59 perfect p p p p p er er er er er er er er er er
f f f f f f f f f ix ix ix ix ix kcl kcl kcl kcl
kcl kcl kcl kcl kcl kcl t h# h# h# h# h# h# h# h# h#

60 74 he hh hh hh hh hh hh hh hh iy iy iy iy iy iy iy

75 152 thought th th th th th th th th ao ao ao ao ao
ao ao ao ao ao ao ao ao ao ao ao ao ao ao ao ao tcl
ch h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h#
h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h#
h# h# h# h# h# h# h# h# h# h# h# h#
```

The first two numbers for each word represent the beginning and ending frame, and the repeated values are the phonetic labels for each 20 msec (overlapping by 10 msecs) frame. Note that for this particular example, the middle sentence pause is recognized as `/h#/`

The ROS Info for Correct Word Transcription Method		
	With Mid-silences	Without Mid-silences
Num. Phones	12	12
Duration (secs)	0.87	0.87
ROS_{IMD}	13.79	13.79
ROS_{MR}	34.52	34.52

Table 2: The ROS for sentence mtcs08-si1972 from the TIMIT training set, using the correct word transcription.

and not /pau/, therefore they will be excluded when calculating the ROS, explaining why the values in the columns “With Mid-silences” and “Without Mid-silences” are identical. The ROS values for our example sentence are reported in Table 2. We notice that for this sentence the ROS_{MR} is over twice as large as the ROS_{IMD}. Upon closer inspection, we see some phone labels with duration of one frame (in the word “thought”, /tcl/ for example) which have a high instantaneous phone ROS, and will boost the overall ROS of the sentence. Also, comparing the results in Tables 1 and 2, we note that the ROS_{IMD} measured using both methods are similar, while the ROS_{MR} calculated using the correct word method is over twice as large as the ROS_{MR} measured using the phonetic hand-labeling.

Using the alignment data, we calculated the ROS for all TIMIT sentences.

5.1.3 Calculating ROS from Hypothesized Word Transcription

The method of calculating ROS using hypothesized word transcriptions is very similar to the one discussed in Section 5.1.2, except that we need to perform a word recognition in order to obtain the hypothesized word sequence. Our example sentence was recognized as “perfect results” by our system, and the phonetic alignment we get from the forced Viterbi alignment follows:

```
10 66 perfect p p p p p er er er er er er er er er er
f f f f f f f f f ix ix ix ix ix kcl kcl kcl kcl kcl
kcl kcl kcl kcl kcl kcl t h# h# h# h# h# h# h# h# h#
h# h# h# h# h# h# h#

67 152 results r r iy iy iy iy iy iy z z z z z z z z
z ah ah ah ah ah ah ah ah ah ah l l l l l l l l l tcl
tcl tcl tcl tcl tcl tcl tcl tcl tcl tcl tcl s h# h#
h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h#
h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h# h#
```

One caveat of this method is that the quality of the results depends on the word recognition accuracy and the complexity of the task. This is a problem particularly for TIMIT

ROS Info for Hypothesized Word Transcription Method		
	With Mid-silences	Without Mid-silences
Num. Phones	13	13
Duration (secs)	0.90	0.90
ROS_{IMD}	14.44	14.44
ROS_{MR}	29.03	29.03

Table 3: The ROS for sentence mtcs08-si1972 from the TIMIT training set, using the hypothesized word transcription.

because it is a difficult word recognition task and the results are generally poor. For example, using a simple word-pair grammar that only includes the training set words, our word recognizer has about 80% error[4]. If we instead use a word-pair grammar derived from the TIMIT training *and test* set (a.k.a. cheating grammar), the word recognition error for the 1344 test sentences is about 26%. A more sophisticated grammar, with probabilities derived from a much larger database would probably improve the word recognition accuracy and thereby the accuracy of the measured ROS.

Note that for our example sentence in Table 3, the ROS_{IMD} measure varied about 45%, while the ROS_{MR} varied about 126% compared to the ROSs calculated using the phonetic hand transcriptions.

5.1.4 Calculating ROS from Hypothesized Phonetic Transcription

Since our decoder did not explicitly provide state path information, we performed a phone recognition on TIMIT and used the hypothesized phone transcriptions for forced alignment. For phone recognition, we used single state models, and the transition probabilities were calculated to match the average phone durations.

The phonetic hand transcription of our example sentence is
`h# p er f ix kcl t pau hh iy th ao tcl h#`
and it is recognized as
`h# p er f axr kcl t pau hh iy s ao tcl h#`.
The phonetic alignment resulting from the forced Viterbi procedure is as follows:

```
0    8 h#       h# h# h# h# h# h# h# h# h# h#
9   14 p        p p p p p p
15  24 er       er er er er er er er er er er
25  33 f        f f f f f f f f f
34  38 axr      axr axr axr axr axr
39  48 kcl      kcl kcl kcl kcl kcl kcl kcl kcl kcl
                kcl kcl
49  55 t        t t t t t t t
56  65 pau      pau pau pau pau pau pau pau pau pau
```

[4] This grammar was used for recognizing our example sentence mtcs08-si1972.

ROS Info for Hypothesized Phone Transcription Method		
	With Mid-silences	Without Mid-silences
Num. Phones	12	11
Duration (secs)	1.06	0.96
ROS_{IMD}	11.32	15.09
ROS_{MR}	11.46	15.55

Table 4: The ROS for sentence mtcs08-si1972 from the TIMIT training set, using the hypothesized phone transcription.

```
                 pau pau
66  67 hh        hh hh hh
68  74 iy        iy iy iy iy iy iy iy iy
75  83 s         s s s s s s s s s s
84  102 ao       ao ao ao ao ao ao ao ao ao ao ao ao
                 ao ao ao ao ao ao ao ao
103 114 tcl      tcl tcl tcl tcl tcl tcl tcl tcl tcl
                 tcl tcl tcl tcl
115 152 h#       h# h# h# h# h# h# h# h# h# h# h# h#
                 h# h# h# h# h# h# h# h# h# h# h# h#
                 h# h# h# h# h# h# h# h# h# h# h# h#
                 h# h# h#
```

Table 4 shows the ROS for each condition. Note that since the mid-sentence pause was identified correctly as /pau/ and not as /h#/, the ROSs for "with mid-silence" and "without mid-silence" condition are different. At least for this example, ROS_{MR} is very different than the one calculated using the correct word transcription.

The phone recognition error on all 5040 sentences of TIMIT is 28.7%, and the error for the 1344 test sentences is 31.8%, and for the core NIST test set (last 200) is 34.0%.

5.2 Observations

As we mentioned earlier, we are looking for an ROS measure that can be reliably and accurately estimated. First, we calculated the ROS using the phonetic hand segmentation, and defined it as the "correct" ROS. Then, we calculated the ROS using the methods discussed above and estimate the "goodness" of the ROS measure by its correlation with the "correct" measure.

5.2.1 Distributions

We plotted the histograms of the distributions of the hand-transcribed ROS measure. Our first observation is that the distribution represents a Gaussian distribution very well. For $ROS_{phn-corr-IMD}$ 68.3% and 95.6%, and for $ROS_{phn-corr-MR}$ 70.2% and 95.0% of the data lies between one and two standard deviations from the mean, respectively. As we see in Figure 2, the ROS distribution resulting from the MR formula has a larger mean and standard deviation compared to the one calculated using the IMD formula. The reason for this difference, as we discussed in Section 5.1.2, is that very short phones have a high instantaneous phone rate, which boosts the overall ROS of the sentence using MR.

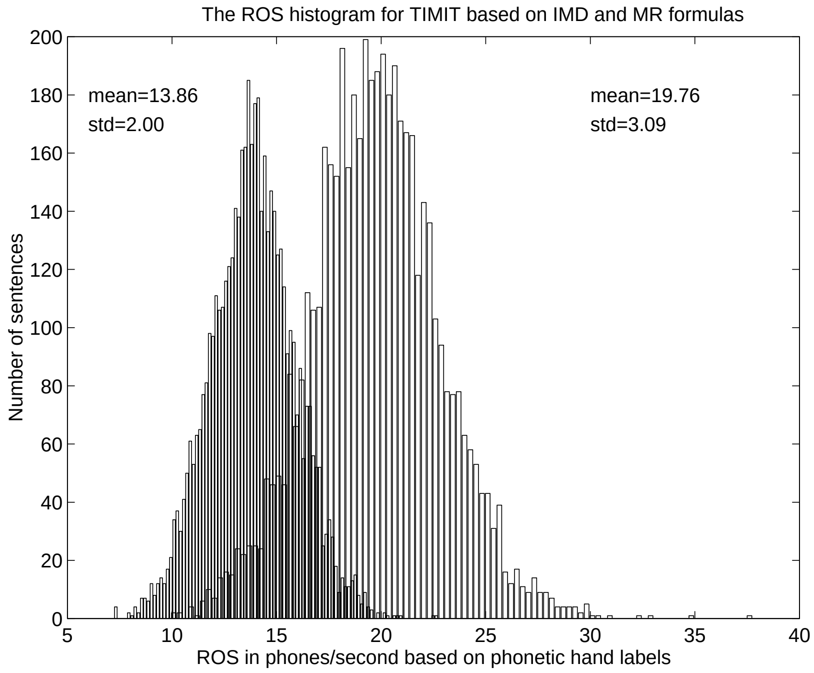


Figure 2: Histogram of rate of speech for TIMIT sentences. The values in the curve on the left have been measured using the IMD formula and the ones on the right using the MR formula.

We also plotted the distribution of the ROS for the male and female sentences separately (see Figures 3 and 4). We see that the mean ROS for male sentences is 2.76% higher than the mean ROS for female sentences. This difference is significant on the $p < 0.001$ level[5]. This rate difference has been previously observed in a study of American vowels duration by Hillenbrand et al. [9]:

> "The pattern of durational differences among the vowels is very similar to that observed in connected speech. Our vowel durations from /hVd/ syllables are two-thirds longer than those measured in connected speech by Black (1949), but correlate strongly (r=0.91) with the connected speech data. There were significant differences in vowel duration across the three talker groups (F[2,33]=9.04, p<0.001). Neuman-Kewls post-hoc analyses showed significantly shorter durations for the men when compared to either the women or the children. Longer durations for the children were expected based on numerous developmental

[5] Although it seems that on average males speak faster than females, it is debatable whether the amount of information content transferred per second is any higher.

> studies (e.g., Smith, 1978; Kent and Forner, 1980) but the differences between the men and the women were not expected. *We do not have an explanation for this finding and do not know if these male-female duration differences would also be seen in conversational speech samples.*"

We also looked at the intra-speaker ROS variability. Based on the phonetic hand transcriptions and the IMD formula, we have measured the mean ROS for the 630 TIMIT speakers to range between [9.56, 17.73] phones/sec, standard deviation [0.44, 3.17] phones/sec, and the coefficient of variation (where coefficient of variation is the standard deviation divided by the mean) ranges between [3.29%, 22.87%].

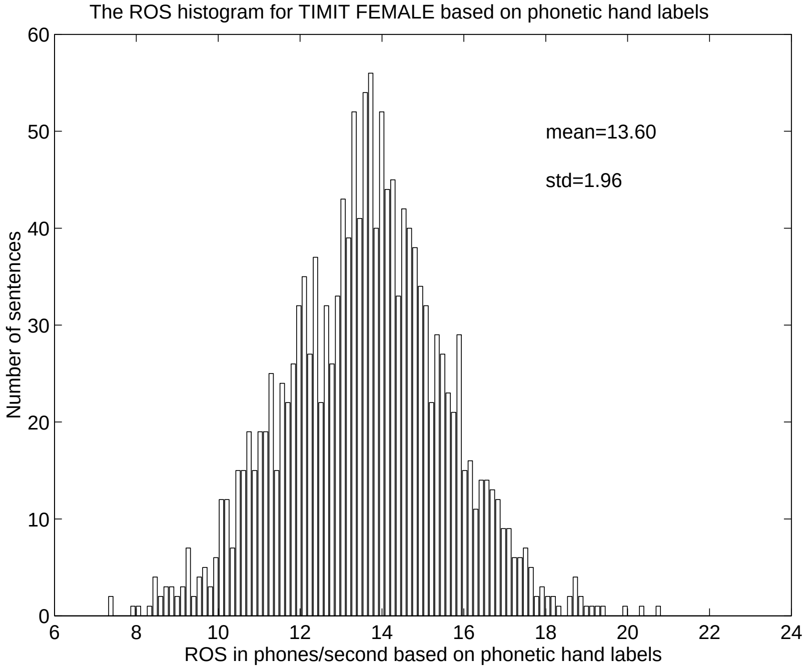


Figure 3: Histogram of rate of speech for TIMIT female sentences, based on the phonetic hand transcriptions and the IMD formula.

For completeness, we include similar plots from the WSJ database (see Figures 5, 6, and 7). The average ROS_{IMD} for the male speakers in WSJ0 is 4.02% higher than the average for the female speakers.

5.2.2 Correlation of the ROS Measures

We generated correlation coefficients between each of the ROS measurement methods and the ROS hand-labeled values. The relevant values are shown in Table 5. We have only included the comparisons with the change of a single experimental variable; for example, it is justified to compare $ROS_{phn-hyp-IMD}$ with $ROS_{phn-corr-IMD}$, and not with $ROS_{phn-corr-MR}$. Even for the phonetically hand-transcribed data, the correlation between

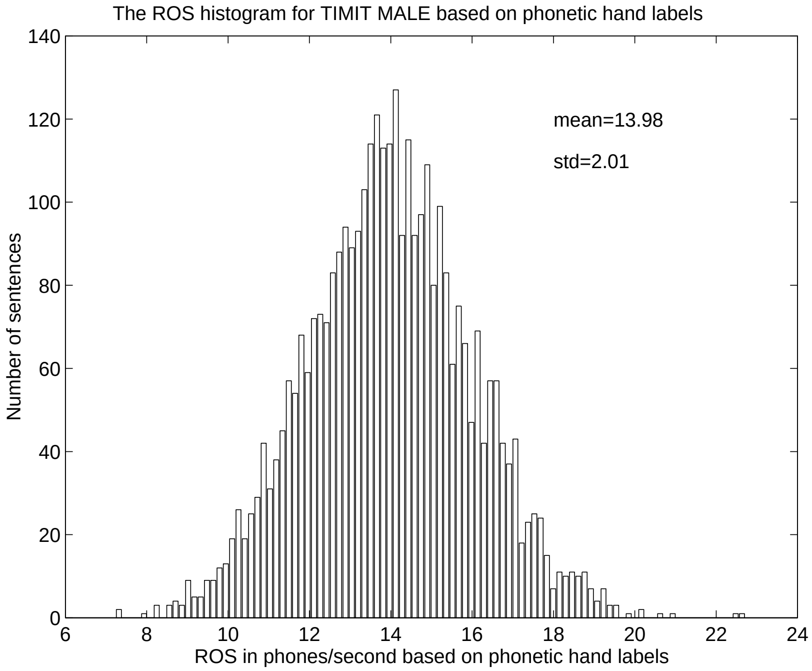


Figure 4: Histogram of rate of speech for TIMIT male sentences, based on the phonetic hand transcriptions and the IMD formula.

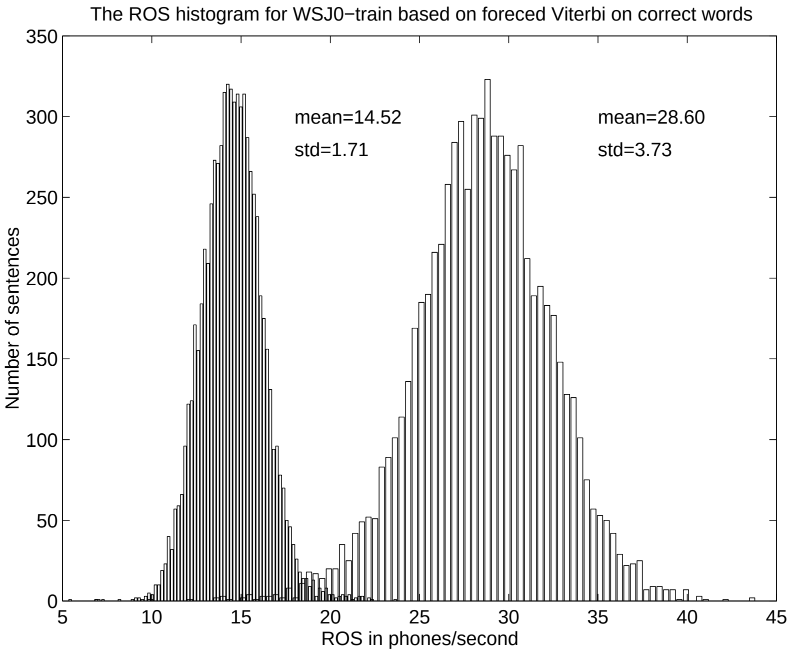


Figure 5: Histogram of rate of speech for WSJ0 training sentences. The values in the curve on the left have been measured using the IMD formula and the ones on the right using the MR formula.

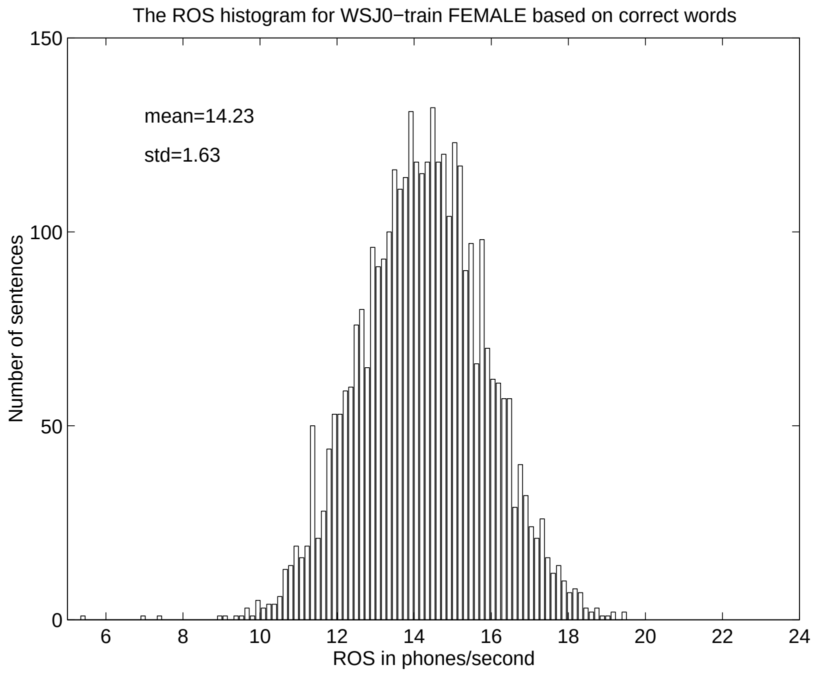


Figure 6: Histogram of rate of speech for WSJ0 training female sentences, based on the correct word transcriptions and the IMD formula.

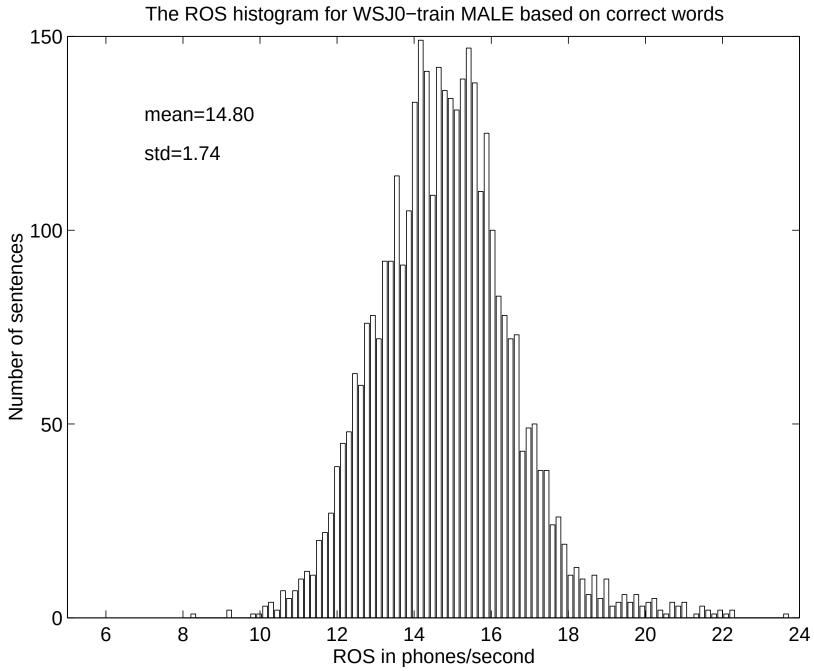


Figure 7: Histogram of rate of speech for WSJ0 training male sentences, based on the correct word transcriptions and the IMD formula.

the ROSs calculated using the IMD and MR formulas is not very high, for example, for TIMIT $\rho(ROS_{phn-corr-IMD}, ROS_{phn-corr-MR}) = 0.77$.

Corr. Coeff. of Different ROS Measures with the Phonetically Hand Transcribed ROS				
	IMD formula		MR Formula	
ROS Method	W/O Mid-sil	W Mid-sil	W/O Mid-sil	W Mid-sil
Wrd Correct	0.88	0.87	0.40	0.40
Phn hypothesized	0.84	0.83	0.61	0.60

Table 5: Correlation coefficients for the 1344 TIMIT test sentences between various methods of calculating the ROS with the phonetically hand transcribed calculated ROS.

As we see in Table 5, the ROSs measured using the MR formula are consistently less correlated with the phonetically hand transcribed ROS. The IMD formula seems to be a more reliable way of estimating the ROS of a sentence. Also, taking out the middle silence seems to make the ROS estimation slightly more consistent. Using the correct word transcription method seems to be superior to using hypothesized phone transcriptions for the IMD formula, and the inverse is true for the MR formula. Perhaps using lexical models derived from the data results in a phonetic segmentation with some short phones, which strongly affect the MR and not the IMD measure. We will come back to this point in Section 8.

Note that we have not used the hypothesized word transcription method for ROS calculation, because we think that this method is particularly unsuitable for TIMIT. TIMIT is primarily a phone recognition task since the test sentences tend to have many previously unseen word pairs, for which we have no language model information. We will revisit these methods for WSJ0 in Section 7.1.

5.3 A Final Note on ROS Measurement

As discussed in the sections above, ROS is not an absolute measure and depends on the calculation method. Furthermore, the size of the phone-set and the alignment procedure can cause differences in the measured ROS. As an illustration of this point, consider sentence 011c0201 from the WSJ0 training set, with the word transcription “The sale of the hotels is part of holiday’s strategy to sell off assets and concentrate on property management.” Table 6 shows the segmentation and labeling for this sentence using ICSI’s HMM/MLP hybrid system and CMU’s SPHINX-II recognition system [10][6]. If we employ the IMD formula for calculating ROS, using ICSI’s segmentation we get 16.94 phones/second and using CMU’s segmentation we get 14.21 phones/sec. One main reason for this difference is that there are more phonetic labels used in the ICSI alignment. Specifically, for each stop consonant, the ICSI alignment is based on two labels: one for the closure and one for the stop phone (see column 11 of Table 6, for example), resulting in 10 more closure phone-labels for the ICSI alignment.

[6] Thanks to Matt Siegler for the raw CMU data.

Order	ICSI's Alignment		CMU's Alignment	
	Phone	Frames	Phone	Frames
0	DH	2	DH	6
1	IH	7	AX	6
2	S	11	S	12
3	EY	7	EY	8
4	L	10	L	10
5	EL	6	AX	4
6	V	4	V	6
7	DH	2	DH	3
8	AX	10	AX	8
9	HH	3	HH	4
10	OW	9	OW	9
11	TCL	5	T	11
12	T	8	EH	8
13	EH	6	L	20
14	L	22	Z	14
15	Z	11	IX	6
16	H#	5	Z	9
17	IX	5	P	9
18	Z	6	AA	6
19	PCL	5	R	3
20	P	5	TD	3
21	AA	6	AX	4
22	R	3	V	6
23	TCL	3	HH	8
24	AXR	4	AA	4
25	V	7	L	9
26	HH	6	AX	3
27	AA	6	D	5
28	L	5	EY	7
29	AX	6	Z	6
30	DCL	3	S	6
31	D	1	T	6
32	EY	10	R	3
33	Z	5	AE	10
34	S	5	DX	3
35	TCL	3	AX	5
36	T	3	JH	9
37	R	6	IY	13
38	AE	8	T	6
39	DX	1	AX	5
40	IH	4	S	11
41	DCL	4	EH	5
42	JH	8	L	5
43	IY	11	AO	10
44	TCL	5	F	6
45	T	2	AE	19
46	IX	5	S	12
47	S	11	EH	4
48	EH	6	TS	17
49	L	8	AX	4
50	AO	8	N	5
51	F	6	DD	4
52	H#	6	K	7
53	AE	13	AA	9
54	S	11	N	3
55	AX	4	S	7
56	TCL	8	AX	3
57	S	6	N	3
58	HH	4	T	10
59	IX	3	R	4
60	N	5	EY	8
61	KCL	4	TD	4
62	K	8	AO	8
63	AA	9	N	4
64	N	2	P	11
65	S	7	R	3
66	AX	4	AA	6
67	N	2	P	6
68	TCL	2	AXR	7
69	T	9	DX	4
70	R	4	IY	9
71	EY	6	M	6
72	TCL	6	AE	8
73	AH	7	N	3
74	N	3	IX	9
75	PCL	5	JH	9
76	P	7	M	6
77	R	2	AX	8
78	AA	7	N	5
79	PCL	5	TD	8
80	P	1	SILE	34
81	AXR	6		
82	DCL	3		
83	D	1		
84	IY	9		
85	M	6		
86	AE	9		
87	N	3		
88	IX	7		
89	JH	12		
90	M	6		
91	IX	3		
92	N	5		
93	TCL	7		
94	H#	33		

Table 6: The automatic alignment for sentence 011c0201 from the WSJ0 training set for ICSI's HMM/MLP hybrid system and CMU's SPHINX-II recognition system.

Note that neither alignment is "more correct" for ROS calculation; we simply wish to demonstrate that ROS is a phone-set dependent measure.

6 Analysis of Fast Speech

From psycho-acoustic experiments [22] we know that when the speaking rate becomes too fast, the production of speech sounds changes, the duration of sounds and syllables generally becomes very short, or phones get omitted altogether. We have considered two reasons for the higher error rate of faster speakers. First, due to increased coarticulation effects, the spectral features of fast speech may be inherently different from normal speech, and if so, these differences must be reflected in the extracted features (*acoustic-phonetic* causes). *Phonological* causes are the second potential culprit: the normal word models may be unsuitable for fast speech because of phonemic durational mismatches (*durational errors*) or phone omission (*deletion errors*). In the following sections, we describe our investigation of these two hypotheses using the TIMIT and the WSJ corpora.

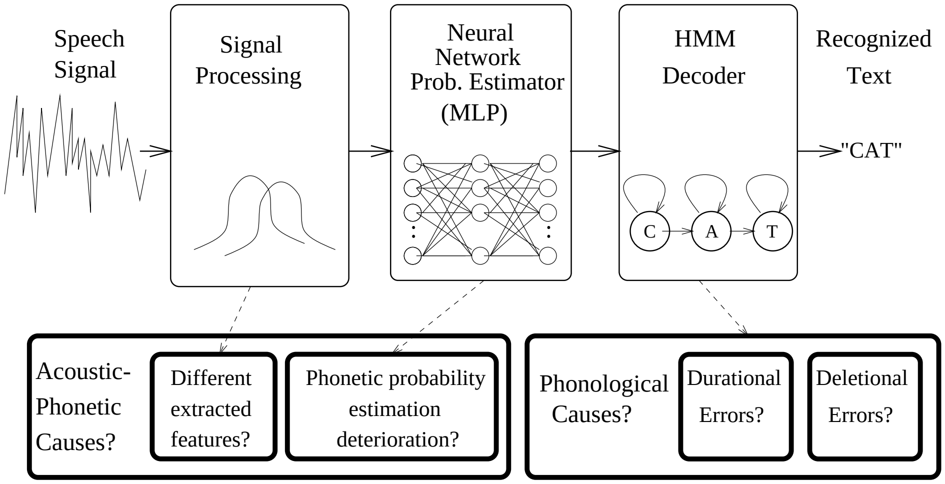

Speech Signal
Signal Processing
Neural Network Prob. Estimator (MLP)
HMM Decoder
Recognized Text
"CAT"
C
A
T
Acoustic-Phonetic Causes?
Different extracted features?
Phonetic probability estimation deterioration?
Phonological Causes?
Durational Errors?
Deletional Errors?

Figure 8: Potential causes of error in fast speech.

6.1 Are the Spectral Features Different?

If shorter phoneme durations increase coarticulation effects, the spectral characteristics must be different for each sound, and the difference should be reflected in the extracted features. Therefore, we should be able to train a classifier to distinguish between *fast* and *slow* phones based on the extracted features. This form of non-parametric hypothesis testing can be useful for such multi-dimensional investigations.

In order to eliminate any word model effects (due to automatic labeling and alignment), we used the hand-labeled TIMIT database and chose 400 sentences from the SI & SX training sentences, 100 for each combination of $\{fastest, slowest\} * \{male, female\}$. Then we calculated the PLP12 & energy features and their deltas [8] (a total of 26 features) for each 20 msec window of speech, overlapped every 10 msec. We trained a two-layer neural network (26 input, 50 hidden, and 2 output units) using the back-propagation algorithm and softmax error criterion for each phone on fast and slow speakers' extracted features. To eliminate gender variabilities, we trained one classifier on female and one on male speakers for each phone. We exploited our limited data using a jack-knifing approach, by training on 90% of the data and testing on the remaining 10% for each of the ten possible splits. For each split we reported the average classification accuracy on all the the holdout test frames.

We calculated the mean classification accuracy for each phone, averaged over both genders and the 10 jack-knifed test scores (see Table 7). The overall mean classification accuracy, averaged over all phones, was 73% (significantly higher than chance) for a total of 120K frames of data. For some phones, such as /uw/, /uh/, /en/, /oy/, /aw/, /ux/, /y/, /ao/, /ow/, /hh/, and /ay/ (mostly diphthongs and glides) the classification score was between 80-90%. This difference makes sense especially in the light of psycho-acoustical studies that suggest diphthongs and glides are most affected by ROS variations [14]. The most difficult phones for speed discrimination were, unsurprisingly, the silence phones, closures, stops, and some fricatives. The training criteria and the architecture of the net could have probably been changed to optimize the discrimination accuracy, but since our objective was only to see that such discrimination was possible, such tunings were not performed.

We then conducted another two sets of net trainings: one only using PLP and energy features (without deltas) to see whether the speed discrimination was possible without use of any dynamic information, and another one only with delta PLP and energy features, to see if the discrimination was possible from the dynamic information alone. Again, we trained two-layer neural networks with 13 input, 93 hidden[7], and 2 output units on each phone to discriminate between fast and slow sentences. The mean classification accuracy averaged over all phones for the *features only* condition was about 72% and for the *deltas only* experiment was about 67% (for details see Tables 8 and 9). From these experiments we conclude that delta information is not necessary for discrimination between individual frames of fast and slow speech for particular phones, and furthermore, the discrimination is more difficult when using only delta (dynamic) information. Perhaps the differences between fast and slow speech frames primarily lie in the differences in the steady state information.

It is evident that features for fast and slow sounds are different. The next question is whether this difference is causing the higher recognition error rate for fast speakers. We tested this hypothesis by examining the error of the MLP phonetic probability estimator for each frame. In order to see this general trend between ROS and the errors of the MLP better, we grouped the sentences in ROS bins with size σ_{ROS}, and boundaries $[\mu_{ROS} + n\sigma_{ROS}, \mu_{ROS} + (n+1)\sigma_{ROS}]$, and calculated the average frame error for each bin (see Figure 9). We see that for sentences which lie outside $\mu_{ROS} \pm \sigma_{ROS}$, the frame error is at least 2 absolute percentage points, or 6 relative percentage points higher.

[7] To keep the number of parameters roughly the same, we increased the number of hidden units from 50 to 93 to compensate for the decrease of input units from 26 to 13.

PHONE	BROAD CAT.	Percent Corr. Discr.	Number Of Frames
h#	sil	55.9	7451
tcl	sil	61.5	1577
kcl	sil	62.3	1316
dcl	sil	62.9	1093
pcl	sil	63.0	932
s	fric	63.2	3427
k	stop	63.3	1055
z	fric	63.4	1534
bcl	sil	65.2	707
d	stop	65.3	319
t	stop	65.3	973
ix	vowel	65.5	1879
f	fric	66.0	1120
g	stop	66.3	210
p	stop	67.1	567
th	fric	67.4	439
dh	fric	67.6	517
l	liq	67.7	1616
gcl	sil	68.0	425
n	nasal	68.5	1816
sh	fric	70.2	838
b	stop	70.6	209
r	liq	70.8	1376
ax	vowel	70.8	864
iy	vowel	70.8	2296
ih	vowel	71.1	1787
w	liq	71.8	737
q	sil	72.9	894
dx	stop	73.4	232
v	fric	74.4	619
eng	nasal	75.0	9 *
m	nasal	75.7	1202
ch	fric	75.8	337
ae	vowel	76.3	1789
ah	vowel	76.4	1166
eh	vowel	76.5	1637
aa	vowel	76.9	1652
epi	sil	77.0	205
jh	fric	77.2	351
ey	vowel	77.2	1470
pau	sil	77.2	1075
ng	nasal	77.5	356
er	vowel	77.9	1103
hv	fric	78.6	277
axr	vowel	78.7	1015
ay	vowel	79.2	1508
el	liq	80.1	416
hh	fric	80.7	292
ow	vowel	81.3	1243
ao	vowel	81.7	1300
y	liq	83.2	309
ux	vowel	84.5	677
aw	vowel	84.9	665
oy	vowel	85.4	305
en	nasal	88.2	210
uh	vowel	88.4	224
ax-h	vowel	89.6	48 *
zh	fric	90.0	87 *
uw	vowel	90.1	213
nx	nasal	91.6	94 *
em	nasal	100.0	31 *

Table 7: Discrimination scores for fast vs. slow phones of TIMIT training set using both PLP and energy and their deltas as features. Each discrimination score is an average of the 10 jack-knifing experiments for males and females. The * next to some columns means there were less than 100 frames of data for these phones and the results are deemed unreliable.

PHONE	BROAD CAT.	Percent Corr. Discr.	Number Of Frames
h#	sil	55.4	7451
dcl	sil	59.7	1093
t	stop	60.4	973
kcl	sil	60.5	1316
k	stop	60.6	1055
tcl	sil	60.6	1577
pcl	sil	61.3	932
d	stop	61.4	319
z	fric	61.6	1534
f	fric	61.8	1120
s	fric	62.0	3427
bcl	sil	62.7	707
ix	vowel	62.9	1879
p	stop	63.2	567
g	stop	63.8	210
gcl	sil	64.4	425
l	liq	65.8	1616
sh	fric	66.0	838
q	sil	66.0	894
th	fric	66.2	439
dh	fric	66.2	517
n	nasal	67.2	1816
r	liq	67.3	1376
ax	vowel	68.1	864
dx	stop	68.5	232
v	fric	68.7	619
ih	vowel	68.9	1787
iy	vowel	69.7	2296
b	stop	69.8	209
w	liq	70.9	737
epi	sil	70.9	205
ch	fric	71.5	337
jh	fric	72.4	351
ng	nasal	72.8	356
ey	vowel	73.2	1470
pau	sil	73.6	1075
axr	vowel	73.6	1015
er	vowel	73.9	1103
ah	vowel	73.9	1166
hv	fric	74.5	277
m	nasal	74.9	1202
eng	nasal	75.0	9 *
hh	fric	75.6	292
eh	vowel	76.1	1637
el	liq	77.1	416
aa	vowel	77.4	1652
ae	vowel	78.3	1789
ay	vowel	78.8	1508
ao	vowel	79.2	1300
y	liq	79.8	309
ow	vowel	80.0	1243
oy	vowel	83.4	305
en	nasal	83.8	210
ux	vowel	83.8	677
aw	vowel	84.4	665
zh	fric	85.4	87 *
ax-h	vowel	85.6	48 *
nx	nasal	87.0	94 *
uh	vowel	88.0	224
uw	vowel	88.8	213
em	nasal	100.0	31 *

Table 8: Discrimination scores for fast vs. slow phones of TIMIT training set using only PLP and energy (without delta) features. Each discrimination score is an average of the 10 jack-knifing experiments for males and females. The * next to some columns means there were less than 100 frames of data for these phones and the results are deemed unreliable.

PHONE	BROAD CAT.	Percent Corr. Discr.	Number Of Frames
h#	sil	54.6	7451
ix	vowel	58.6	1879
s	fric	59.5	3427
ih	vowel	59.7	1787
r	liq	60.4	1376
z	fric	60.7	1534
ao	vowel	60.8	1300
l	liq	61.0	1616
w	liq	61.1	737
ax	vowel	61.1	864
ae	vowel	61.2	1789
iy	vowel	61.3	2296
dcl	sil	61.5	1093
tcl	sil	61.6	1577
aa	vowel	61.6	1652
k	stop	61.9	1055
g	stop	62.1	210
pcl	sil	62.4	932
n	nasal	62.5	1816
er	vowel	62.6	1103
sh	fric	62.9	838
kcl	sil	63.2	1316
m	nasal	63.3	1202
t	stop	63.5	973
ey	vowel	64.0	1470
gcl	sil	64.2	425
eh	vowel	64.3	1637
ay	vowel	64.7	1508
ah	vowel	64.8	1166
bcl	sil	65.0	707
f	fric	65.0	1120
q	sil	65.3	894
axr	vowel	65.4	1015
th	fric	65.4	439
epi	sil	65.7	205
d	stop	65.9	319
oy	vowel	66.1	305
dh	fric	66.1	517
v	fric	66.5	619
ng	nasal	67.5	356
b	stop	68.0	209
pau	sil	68.6	1075
ow	vowel	68.8	1243
jh	fric	68.8	351
aw	vowel	69.0	665
p	stop	69.1	567
el	liq	70.3	416
en	nasal	70.6	210
dx	stop	71.0	232
ux	vowel	71.3	677
hv	fric	71.5	277
ch	fric	72.8	337
zh	fric	73.3	87 *
hh	fric	73.6	292
y	liq	74.4	309
eng	nasal	75.0	9 *
uw	vowel	75.2	213
uh	vowel	76.3	224
nx	nasal	78.5	94 *
ax-h	vowel	78.8	48 *
em	nasal	97.5	31 *

Table 9: Discrimination scores for fast vs. slow phones of TIMIT training set using delta PLP and energy (without PLP and energy features themselves) as features. Each discrimination score is an average of the 10 jack-knifing experiments for males and females. The * next to some columns means there were less than 100 frames of data for these phones and the results are deemed unreliable.

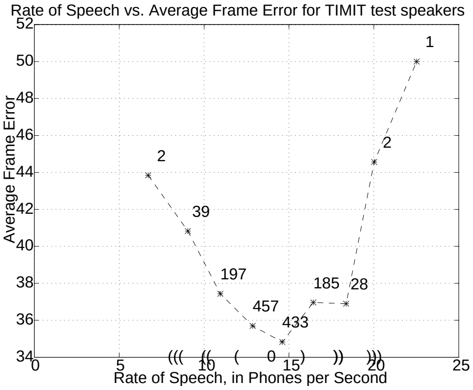


Figure 9: Rate of speech vs. MLP frame error for TIMIT test sentences. Each point represents the average error for a given ROS bin. The numbers on the graph denote the number of sentences in each bin.

6.2 A Closer Look at the Word Models

The next question is whether the higher error rate is due to a mismatch with the word models. One hypothesis is that the implicit durational models in our recognizer do not match the durations used by fast speakers. We have observed that fast speakers tend to favor shorter phone durations and violate phonemic minimum duration requirements (*durational errors*), and also omit phones in their pronunciations altogether (*deletion errors*).

We transcribed a total of 25 sentences for five fast speakers in the WSJ-93 development and evaluation sets by hand and compared their pronunciations with what our single-pronunciation word models predict. We aligned each transcribed word with its corresponding word-model phonetic sequence, using dynamic programming with a distance metric based on the number of phonetic features (e.g., consonant, frontness, height) that differ between two phones, producing a deletion error score.

As noted before, our word models (as with many other systems) have a minimum duration constraint, which require that each phone be repeated for at least n states.[8] For the five transcribed speakers, we also calculated a duration error score which represents how often the transcribed phones were shorter than the minimum duration in the word model. We did not observe a strong correlation between ROS and overall alignment error rate. There were, however, weak correlations between ROS and either of duration and deletion errors. When the two error sources were summed, we found a stronger correlation with ROS. This suggests that both unusually short sounds and deleted sounds may be measurable sources of error in our speech recognizer. However, since we had very limited hand transcribed data, we repeated this experiment on the TIMIT database. Similar to the analysis in 6.1, we divided the sentences into ROS bins, each $\frac{1}{2}\sigma_{ROS}$ wide. There was almost no correlation between ROS and deletion errors alone[9] (Figure 10). The correlation between ROS and durational errors was significantly higher at 0.84 (Figure 11). Combining the deletion and

[8] The value of n in our system is calculated as half of the back-off triphone context-dependent average duration of a phone, estimated from the training data.

[9] For calculating the correlation, we disregarded the bins with less than five sentences.

duration errors, the correlation increases to 0.93 (Figure 12).

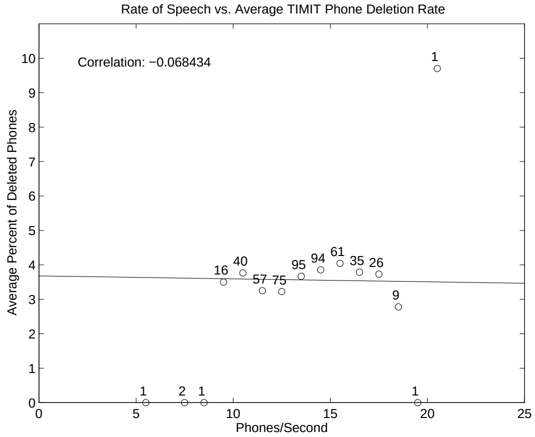


Figure 10: Rate of speech vs. average deletion errors for TIMIT training sentences. The integers on the plot represent the number of sentences in each ROS bin. Bins with less than five sentences were ignored for calculating the correlation coefficient.

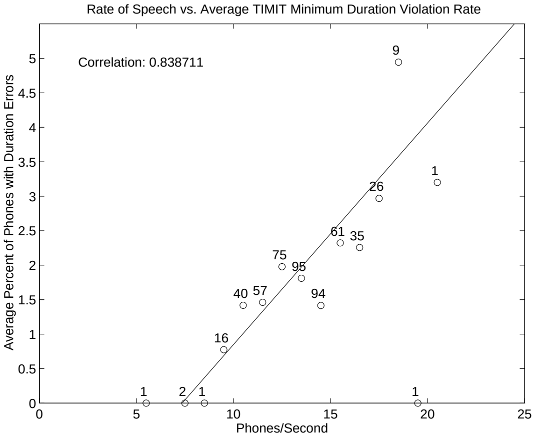


Figure 11: Rate of speech vs. average duration errors for TIMIT training sentences. The integers on the plot represent the number of sentences in each ROS bin. Bins with less than five sentences were ignored for calculating the correlation coefficient.

The effect of deletion errors alone appears to be minor. Perhaps we do not see a consistent correlation between ROS and deletion errors because phone deletion occurs selectively given a particular phone context. This is an argument for applying deletion modeling rules judiciously (Section 7.4).

From these observations we conclude that the *combination* of unusually short sounds and deleted sounds are measurable sources of error in our speech recognizer. We will suggest antidotes in Section 7.3.

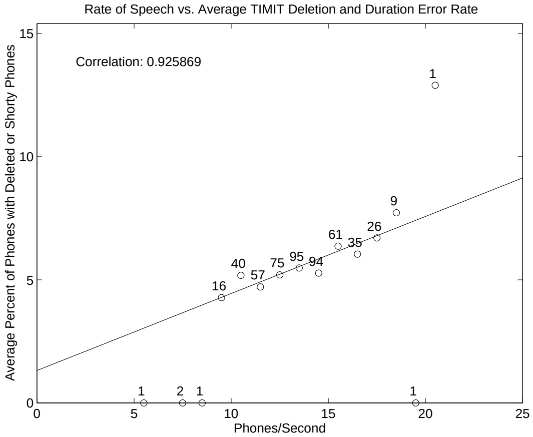


Figure 12: Rate of speech vs. average duration & deletion errors for TIMIT training sentences. The integers on the plot represent the number of sentences in each ROS bin. Bins with less than five sentences were ignored for calculating the correlation coefficient.

7 Some “Antidotes” Against Fast Speech Errors

In the following two sections, we discuss our experiments in trying to alleviate the higher error rates of fast speech. Figure 13 shows the outline of these experiments and Figure 14 shows the over-all structure of our experimental ASR system. All the experiments were run on the WSJ0 corpus, and we have used the WSJ0-93 evaluation set for our tests because two of the ten speakers in this test set are very fast speakers and they provide a good benchmark. Our baseline WSJ0 recognizer is a gender-independent system, with context-independent and one phone per state word models, and utilizes a 5K bigram grammar. It has 16.1% word error for the WSJ0-93 evaluation set.

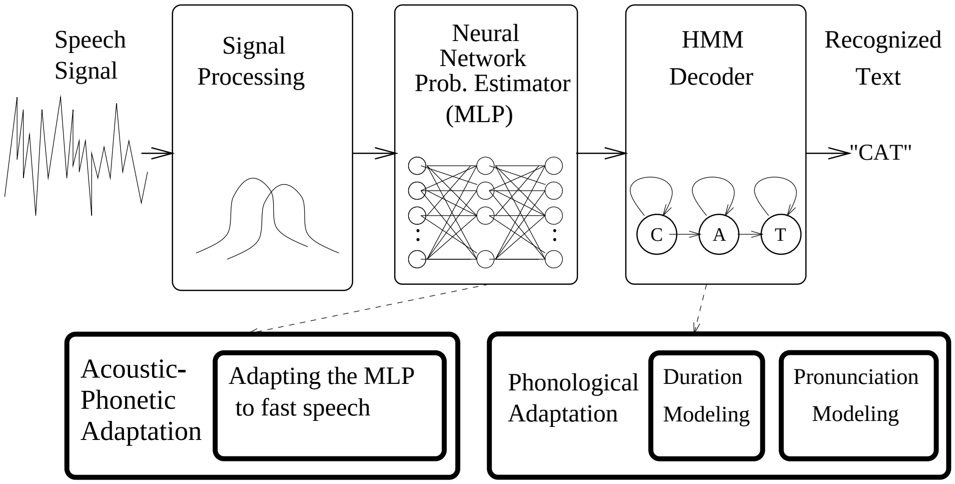


Figure 13: Potential compensations for errors caused by fast speech.

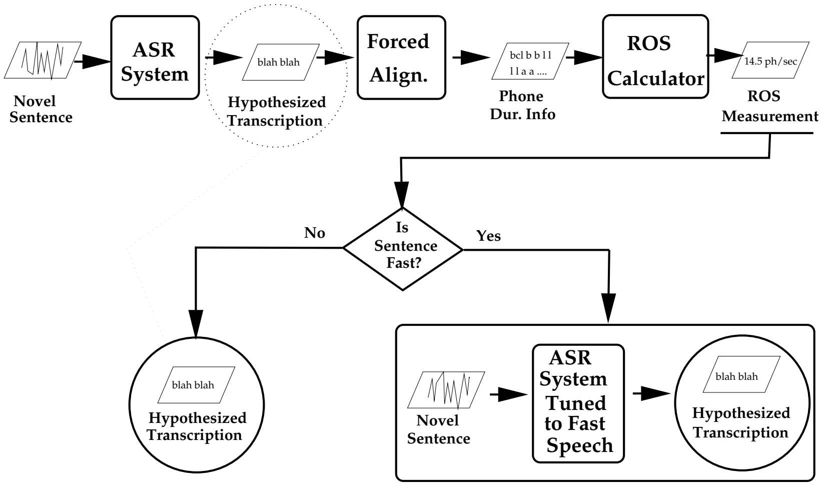


Figure 14: The over-all structure of our rapid-speech-tuned ASR system.

7.1 Using an ROS Estimator

In Section 4.5, we discussed the merits of various ways of calculating the ROS for a sentence without phonetic hand transcription. We concluded (for TIMIT) that in the absence of phonetic hand transcription, using the correct word transcriptions was the best method for calculating ROS, followed by the hypothesized phone transcriptions. Here, we briefly look at how each of the methods operate on the WSJ0-93 evaluation set and choose a set of "fast" sentences.

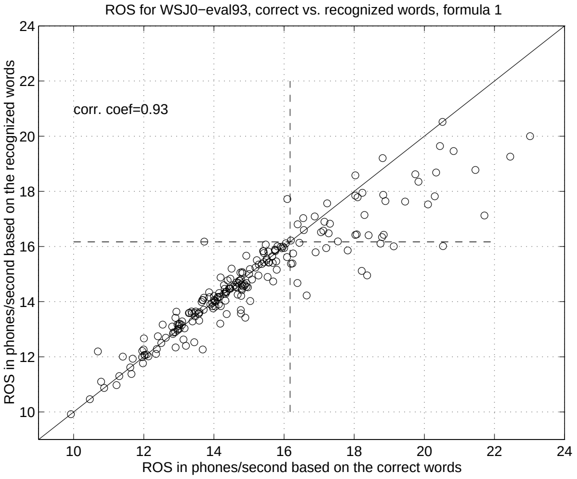


Figure 15: The plot shows the relationship between the correct word transcription method with the hypothesized word transcription for the WSJ0-93 Eval sentences, based on the IMD formula. The dashed lines are drawn at $\mu + 1.65\sigma$.

We see in Figures 16 and 15 that the ROS calculated using the hypothesized word transcriptions has higher correlation with the ROS calculated using the correct word tran-

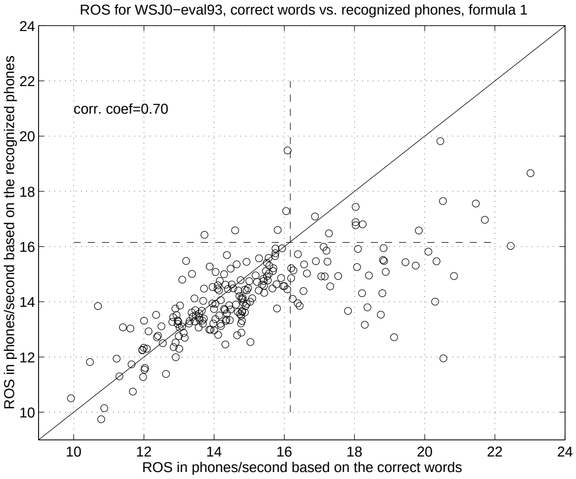


Figure 16: The plot shows the relationship between the correct word transcription method with the hypothesized phone transcription for the WSJ0-93 Eval sentences. The dashed lines are drawn at $\mu + 1.65\sigma$.

scriptions than using the hypothesized phone transcriptions, which is contrary to what we observed for TIMIT in Table 5. Each dot on the Figures 16 and 15 represents one of the sentences in the test set. The sentences to the right of the vertical line (drawn at $\mu + 1.65 * \sigma$) are the sentences deemed fast by the correct word transcription method. The sentences lying above the horizontal line (also drawn at $\mu + 1.65 * \sigma$) are the ones the ROS estimators chose as fast. In each case, the sentences lying in the fourth quadrant are the fast sentences that are missed. We see that more fast sentences are missed for the hypothesized phone case than in the hypothesized word. Perhaps for a tasks where the word recognition accuracy is acceptable, hypothesized words provide a better technique than hypothesized phones for estimating the ROS. This may be because the word models provide a constraint in addition to the acoustic-phonetic information which helps to determine phone boundaries. Yet, a phone recognition pass can be faster and may be an acceptable alternative.

7.2 Acoustic-Phonetic Modeling: Retraining The MLP

Based on our observations in Section 6.1, we decided to adapt our MLP phonetic estimator to fast speech. We chose the fastest 5% of the sentences based on three criteria:

- ROS calculated based on the correct word transcription, without taking out the silence durations (*corr-wrd-Wsil*)
- ROS calculated based on the correct word transcription (*corr-wrd*)
- ROS calculated based on the hypothesized phone transcription (*hyp-phn)*

The first and last experiments, using *corr-wrd-Wsil* and *hyp-phn* as criteria are reported here for completeness. These experiments were run prior to our study of the ROS measures, which showed that these two methods are not optimal. Roughly one third of the sentences chosen by these two criteria and the *corr-wrd* criterion were different.

The ROS cutoff for each case was $\mu + 1.65\sigma$, or 16.17, 17.34, and 16.15 phones/second for each case, respectively. Our 4000 hidden unit MLP was previously trained on all of WSJ0. We adapted this net to the top 5% fast sentences of the training set (360 fast sentences[10]) by retraining the net for three more epochs, at decreasing learning rates of 0.008, 0.004, and 0.002. This is a schedule that we have previously found useful for this kind of retraining.

We tested this adapted net on the WSJ0-93 evaluation set. We looked at the word recognition error rate of fast sentences with $ROS > C$ and slow and medium sentences with $ROS < C$, where C, the cutoff, was either defined to be $\mu + 1.65\sigma$ (Table 10) or $\mu + 1.00\sigma$ (Table 11).

Relative Improvement in Word Error for WSJ-93 Eval Set						
Net Adapt. Crit.	Criteria for Choosing Fast Sentences over $\mu + 1.65\sigma$					
	Correct Word (idealized)		Hyp. Phone		Hyp. Word	
	33 fast	overall	17 fast	overall	21 fast	overall
corr-wrd-Wsil	16.7	5.6	15.5	1.2	10.2	1.9
corr-wrd	2.6	1.2	2.1	0.6	-5.3	-.6
hyp-phn	5.2	1.9	8.8	0.6	4.2	0.6

Table 10: The table shows the percent improvement in recognition word error for the WSJ-93 Evaluation set, after retraining the acoustic probability estimator. Each row shows a different criterion for retraining the MLP (see text for explanation). The column categories correspond to each of the three ways of estimating the ROS during recognition. The first sub-column is the improvement of the fast sentences (which are over the cutoff) relative to the baseline system, and the number in the second sub-column is the percent relative improvement for the overall recognition score (215 sents).

Relative Improvement in Word Error for WSJ-93 Eval Set						
Net Adapt. Crit.	Criteria for Choosing Fast Sentences over $\mu + 1.00\sigma$					
	Correct Word (idealized)		Hyp. Phone		Hyp. Word	
	50 fast	overall	44 fast	overall	37 fast	overall
corr-wrd-Wsil	15.0	6.8	10.9	3.1	14.4	4.3
corr-wrd	5.8	2.5	2.4	0.6	5.9	1.9
hyp-phn	6.5	3.1	9.0	2.5	.9	1.9

Table 11: The table shows the percent improvement in recognition word error for the WSJ-93 Evaluation set, after retraining the acoustic probability estimator. Each row shows a different criterion for retraining the MLP (see text for explanation). The column categories correspond to each of the three ways of estimating the ROS during recognition. The first sub-column is the improvement of the fast sentences (which are over the cutoff) relative to the baseline system, and the number in the second sub-column is the percent relative improvement for the overall recognition score (215 sents).

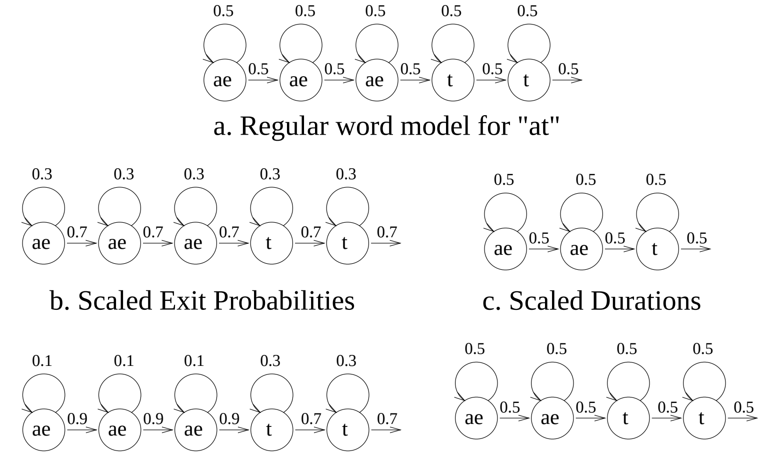
[10] For the first experiment 367 sentences were above the "fast" cutoff and were selected.

From Tables 10 and 11 we conclude that by lowering the ROS cutoff from 1.65σ to 1.00σ and thereby allowing more sentences to benefit from the fast-speech modification, the overall improvement for the test set increases. The second observation is that the *corr-wrd-Wsil* criterion for choosing the fast sentences for adapting the MLP has outperformed the other two criteria. This may be related to the first observation: perhaps choosing the top 5% fastest sentences for training (which corresponds to the sentences with a ROS greater than $1.65\sigma + \mu$) is too restrictive and the threshold should be relaxed. In the case of the *corr-wrd-Wsil* criterion, which does not exclude begin, end, or middle silences, the ROS calculation is not as precise as the other criteria and some "medium fast" sentences may have been used for the adaptation. In any case, it is interesting to note that simply adapting the MLP to fast speech for a few epochs can improve robustness to other fast sentences. We must note that there are many other methods of adapting an MLP that were not explored further. Some of these approaches are discussed in [20, 1] may be used for better adaptation.

A final observation from Tables 10 and 11 is that estimating the ROS of the test sentences using the correct word transcription improved the performance more than using the hypothesized words, and the latter was in turn better than using the hypothesized phones. This is in line with what we had predicted in Section 7.1.

7.3 Duration Modeling

We have investigated methods of adjusting the durational models of phones in order to compensate for ROS effects. Our current phone model, shown in Figure 17.a, requires a minimal duration constraint. For phones that are shorter than the minimum duration, this constraint will sharply decrease the probability of the phone (and consequently, the word which contains the phone) representing the acoustic input.



a. Regular word model for "at"

b. Scaled Exit Probabilities

c. Scaled Durations

d. Graded Scaling of the Probabilities

e. Scaled Vowel Durations

Figure 17: Examples of word models for "at".

In Figure 17.b, we show a model in which we have scaled the probabilities of each HMM state to favor leaving rather than staying in each state. In Table 12 we report cases in which we scaled the exit probability from 0.4 to 0.95. Increasing the exit probability to 0.9 provides the best overall improvement for the test sentences.

Relative Improvement in Word Error for WSJ-93 Eval Set						
Lex. Adapt. Crit.	The Criteria for Choosing Fast Sentences over $\mu + 1.00\sigma$					
	Correct Word (idealized)		Hyp. Phone		Hyp. Word	
	50 fast	overall	44 fast	overall	37 fast	overall
probshift 0.4	-6.8	-2.4	-7.1	-1.8	-7.4	-1.8
probshift 0.6	10.2	4.3	7.1	1.8	8.1	2.4
probshift 0.7	12.5	5.5	6.6	1.8	10.0	3.1
probshift 0.75	12.9	5.5	5.6	1.8	10.0	3.1
probshift 0.8	14.9	6.8	10.9	3.1	12.2	3.7
probshift 0.85	15.3	6.8	9.0	2.4	11.1	3.1
probshift 0.9	17.3	7.4	14.2	3.7	12.2	3.7
probshift 0.95	11.9	5.5	9.9	2.4	7.4	2.4

Table 12: The table shows the percent improvement in recognition word error for the WSJ-93 Evaluation set, after retraining the acoustic probability estimator. Each row shows a different criterion for retraining the MLP (see text for explanation). The column categories correspond to each of the three ways of estimating the ROS during recognition. The first sub-column is the improvement of the fast sentences (which are over the cutoff) relative to the baseline system, and the number in the second sub-column is the percent relative improvement for the overall recognition score (215 sents).

We know that in fast speech, the duration of vowels changes the most, while the duration of stops is relatively constant. Therefore, a variation on the above modification is to **1)** increase the exit probability of only vowels, **2)** increase the exit probability of all phones, except for the stops, and **3)** increase the exit probabilities in a graded scale with stops at the bottom of the scale, vowels on top, and all other phones graded in between. The grading scheme was developed using the knowledge that certain manners of articulation (e.g. vowels) are more likely to shorten in fast speech than others (e.g. stops) [14]. The exit probabilities of each phone were set in an articulation manner dependent fashion; for example, in the 0.7-0.9 lexicon, the assigned probabilities are reported in table 13.

The scale factor is a subjective measure of relative duration change for the particular

Manner	Scaling Factor	Probability
Stops	0.0	0.70
Affricates	0.2	0.74
Fricatives	0.2	0.74
Nasals	0.4	0.80
Liquids	0.7	0.84
Glides	0.7	0.84
Vowels	1.0	0.90

Table 13: The scaling factor in the left column is a subjective measure of relative duration change for a particular manner of articulation; the right column is a mapping from the scaling factor to the probability range [0.7,0.9].

manner of articulation. Although the scale factors have not been optimized, this scaling method shows promise for handling fast speech.

Table 14 shows the results of these variations. The third scheme proved to be the best: increasing the exit probability of the vowels to 0.9 and the stops to 0.7, and the rest of the phones between 0.7 and 0.9.

Relative Improvement in Word Error for WSJ-93 Eval Set						
Lex. Adapt. Crit.	Criteria for Choosing Fast Sentences over $\mu + 1.00\sigma$					
	Correct Word (idealized)		Hyp. Phone		Hyp. Word	
	50 fast	overall	44 fast	overall	37 fast	overall
vowprobshift 0.6	-1.3	-.6	.4	.6	.7	.6
vowprobshift 0.7	6.4	3.1	3.3	1.2	7.0	1.8
vowprobshift 0.8	-8.5	-3.7	-11.3	-2.4	-11.4	-3.1
nostopprobshift 0.7	11.2	4.9	8.5	2.4	9.2	2.4
nostopprobshift 0.8	12.9	5.5	6.6	1.8	10	3.1
nostopprobshift 0.9	18.7	8.0	8.0	1.8	12.5	3.7
gradedshift 0.5-0.7	8.8	3.7	6.6	1.8	8.1	2.4
gradedshift 0.5-0.9	18.0	8.0	13.2	3.1	16.6	4.9
gradedshift 0.7-0.9	24.4	10.5	23.2	5.6	22.6	6.2

Table 14: The table shows the percent improvement in recognition word error for the WSJ-93 Evaluation set, after retraining the acoustic probability estimator. Each row shows a different criterion for retraining the MLP (see text for explanation). The column categories correspond to each of the three ways of estimating the ROS during recognition. The first sub-column is the improvement of the fast sentences (which are over the cutoff) relative to the baseline system, and the number in the second sub-column is the percent relative improvement for the overall recognition score (215 sents).

An alternative would be to simply reduce the minimum phone durations. We tried this in both phone-independent (Figure 17.c) and phone-specific (Figure 17.e) duration scaling experiments. For the phone-independent models, experiments were conducted where 0.5 to 5 frames were subtracted from the average back-off trigram context-dependent duration of each phone. This resulted in an average of 0.25 to 2.5 state deletions in the minimum duration of phone models. The results are reported in Table 15. The best overall results were obtained by deducting 4 frames from the average vowel duration.

The best overall improvements in this section were obtained with the *graded exit probability scaling 0.7-0.9* scheme. Assuming an ideal ROS detector (which knows about the correct word transcription), the relative improvement on the fastest 50 sentences is 24.5% and the relative overall improvement is 10.6%. Using a more realistic ROS detector based on the hypothesized words, the fast and overall relative improvements are 22.6% and 6.2% respectively, and using the hypothesized phones for ROS calculation, the relative improvements are 23.2% and 5.6% respectively.

7.4 Pronunciation Modeling

Relative Improvement in Word Error for WSJ-93 Eval Set						
Lex. Adapt. Crit.	Criteria for Choosing Fast Sentences over $\mu + 1.00\sigma$					
	Correct Word (idealized)		Hyp. Phone		Hyp. Word	
	50 fast	overall	44 fast	overall	37 fast	overall
mindur 0.5	4.0	1.8	3.7	1.2	4.0	1.2
mindur 1.0	5.7	2.4	2.3	.6	5.9	1.8
mindur 2.0	5.1	2.4	3.7	1.2	5.1	1.8
mindur 3.0	2.3	1.2	-5.2	-1.2	.7	.6
vowmindur 1.0	3.0	1.2	3.7	1.2	5.1	1.8
vowmindur 2.0	6.1	2.4	2.3	.6	6.2	1.8
vowmindur 3.0	6.8	3.1	3.3	1.2	7.0	1.8
vowmindur 4.0	7.4	3.1	1.4	.6	6.2	1.8
vowmindur 5.0	3.7	1.8	-.4	0.0	4.8	1.2

Table 15: The table shows the percent improvement in recognition word error for WSJ-93 Evaluation set. Each row shows a different adaptation criterion for the MLP (see text for explanation). The column categories correspond to each of the three ways of estimating the ROS. The first sub-column is improvement of the fast sentences (which are over the cutoff) relative to the baseline system, and the number in the second sub-column is the percent improvement for the overall recognition score.

Finally, we introduced alternate pronunciations into our word models which represent the phone reduction and deletion effects often seen in fast speech [35, 36, 37, 11]. These pronunciations were generated by twenty surface-phonological rules (Table 18) applied to the base (single pronunciation) lexicon. These rules provided an average of 2.41 pronunciations per word for the 5k WSJ test set lexicon. The results of running with this lexicon and the adapted net were insignificantly worse than the base system. However, when performing an error analysis on the results, we noted that the difference in error rate on a sentence-by-sentence basis between the two systems varied widely; for some sentences the base lexicon did much better, and for others, the deletion lexicon removed up to 75% of the errors. It has been reported by other researchers [12, 31] that modifying the word models by using pronunciation rules has not resulted in any improvements for fast speech. One reason may be that these reductions and deletions are more often observed in conversational than read speech. Another possibility is that the rules must be applied judiciously to a subset of words (to *function words*, for example), instead of the whole lexicon. Finally, rules may need to be applied at more limited times, depending, for instance, on more local rate estimates, and on previous words or phones.

7.5 Merging The Solutions

We combined the most promising of the above approaches by using the phonetic probabilities from the adapted net and the ROS-tuned lexicon (Figure 17.b) for decoding (see Table 16).

In some cases, the merged system, i.e., the adapted net combined with the adapted lexicon, outperformed each of the two adaptations alone. For example, using the net

Name	Rule	Example
Reductions		
Syllabic n	[ax ix] n → en	butt**on**
Syllabic m	[ax ix] m → em	bott**om**
Syllabic l	[ax ix] l → el	bott**le**
Syllabic r	[ax ix] r → axr	butt**er**
Flapping	[tcl dcl] [t d]→ dx /V ___ [ax ix axr]	bu**tt**on
Flapping-r	[tcl dcl] [t d]→ dx /V r ___ [ax ix axr]	bar**t**er
H-voicing	hh → hv / [+voice] ___ [+voice]	a**h**ead
L-deletion	l → Ø/ ___ y [ax ix axr]	mi**ll**ion
Gliding	iy → y / ___ [ax ix axr]	colon**i**al
Nasal-deletion	[n m ng] → Ø/ ___ [-voice -consonant]	ra**n**t[11]
Function words		
h-deletion	h → Ø/ # ___	**h**e, **h**im
w-deletion	w → Ø/ # ___	**w**ill, **w**ould
dh-deletion	dh → Ø/ # ___	**th**is, **th**ose
Dental-deletion	[tcl dcl] [t d] → Ø/ [+vowel] ___ [th dh]	brea**d**th
Final dental-deletion	([tcl dcl]) [t d] → Ø/ [+cons +continuant] ___ #	sof**t** (as)
Slur	ax → Ø/ [+consonant] ___ [r l n] [+vowel]	cam**e**ra
Stressed Slur	[+vowel +stress] r → er	w**ar**ts
Pre-stress Contraction	ax → Ø/ [+cons] ___ [+cons] [+vowel +stress]	s**e**nility
Ruh-reduction	r ax → er / [-word bdry] ___ [-word bdry]	sep**ar**able
Transitional Stops[12]		
t-introduction	Ø→ tcl / [+dental +nasal] ___ [+fricative]	prin[**t**]ce
t-deletion	[tcl] → Ø/ [+dental +nasal] ___ [+fricative]	prin**t**s

Figure 18: The table shows the twenty surface-phonological reduction and deletion rules with which we modified our base (single) pronunciation lexicon.

adapted based on the *corr-word-Wsil* ROS criterion and the lexicon with 0.9 probability shift outperforms either of the two schemes. However, the lexicon with the graded probability shift 0.7-0.9, which has proved to be the best modification so far, degrades when combined with the output of the adapted net. Perhaps both modifications are making up for the same fast speech differences, and when combined together, may do “over-modification”. Generally speaking, at this point we have not worked out the ideal combination of these strategies.

8 Discussion

We have discussed methods that result in significant improvements ($p < 0.01$) for fast sentences. Our findings contradict earlier results by Siegler and Stern [31]. We think the

Relative Improvement in Word Error for WSJ-93 Eval Set							
Net Adapt. Crit.	*Lex. Adapt. Crit.*	Criteria for Choosing Fast Sentences over $\mu + 1.00\sigma$					
		Correct Word (idealized)		Hyp. Phone		Hyp. Word	
		50 fast	overall	44 fast	overall	37 fast	overall
corr-wrd-Wsil	gradedshift 0.7-0.9	22.4	9.9	10.9	3.1	19.2	5.5
corr-wrd-Wsil	gradedshift 0.5-0.7	14.6	6.2	6.6	1.8	14.0	3.7
corr-wrd-Wsil	gradedshift 0.5-0.9	21.0	9.3	11.8	3.1	17.4	4.9
corr-wrd-Wsil	probshift 0.8	18.0	8.0	15.6	3.7	15.1	4.3
corr-wrd-Wsil	probshift 0.85	19.3	8.6	15.6	3.7	14.0	3.7
corr-wrd-Wsil	probshift 0.9	18.7	8.0	18.4	4.9	14.0	3.7
corr-wrd	gradedshift 0.7-0.9	13.9	6.2	5.2	1.2	11.1	3.1
corr-wrd	probshift 0.9	9.5	4.3	.1	1.8	7.4	2.4
hyp-phn	gradedshift 0.7-0.9	12.9	5.5	9.0	2.4	12.5	3.7
hyp-phn	probshift 0.9	12.9	5.5	12.3	3.1	13.3	3.7

Table 16: The table shows the percent improvement in recognition word error for WSJ-93 Evaluation set. Each row shows a different adaptation criterion for the MLP (see text for explanation). The column categories correspond to each of the three ways of estimating the ROS. The first sub-column is improvement of the fast sentences (which are over the cutoff) relative to the baseline system, and the number in the second sub-column is the percent improvement for the overall recognition score.

differences in our findings are mostly due to the measure we have used for quantifying ROS. Siegler and Stern use the MR formula, while we have chosen the IMD formula. If we use the MR formula for choosing fast sentences in the WSJ0-93 evaluation set, only thirteen sentences are chosen as fast, with $C = \mu + 1.65\sigma$ (see Figures 19 and 20, compared to Figures 15 and 16). Table 17 shows the effects of our best adaptation technique when either the MR or the IMD measures are used for choosing fast test sentences. We see that the hypothesized phone method for MR is better than the other two measures, which is consistent with the correlation coefficients we observed in Section 5.2.2. Siegler and Stern used the MR measure in combination with the correct and hypothesized word transcription method for calculating ROS, which may partly explain why their improvements were not significant.

Relative Percent Improvement in W.E.R. for WSJ-93 Eval, $C = \mu + 1.00 * \sigma$				
	IMD Measure		MR Measure	
ROS Estimation Criteria	fast	overall (215 sents)	fast	overall (215 sents)
Correct Word (idealized)	24.4% (50 sents)	10.5%	16.1% (13 sents)	2.5%
Hypothesized Word	22.6% (37 sents)	6.2%	22.6% (7 fast)	1.9%
Hypothesized Phone	23.2% (44 sents)	5.6%	11.5% (58 fast)	3.7%

Table 17: The table shows the percent improvement in recognition word error for the WSJ-93 Evaluation set for the IMD vs. the MR measure. The “fast” sub-column is improvement of the fast sentences (which are over the cutoff) relative to the baseline system, and the “overall” sub-column is the percent improvement for the whole test set.

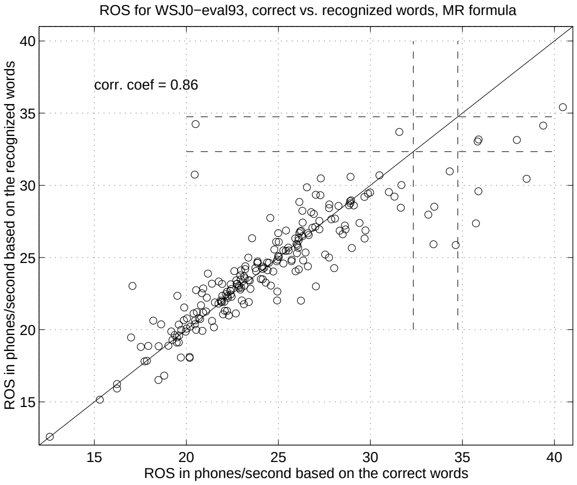


Figure 19: The plot shows the relationship between the correct word transcription method with the hypothesized word transcription for the WSJ0-93 Eval sentences, based on the MR formula. The dashed lines are drawn at $\mu + 1.00\sigma$ and $\mu + 1.65\sigma$.

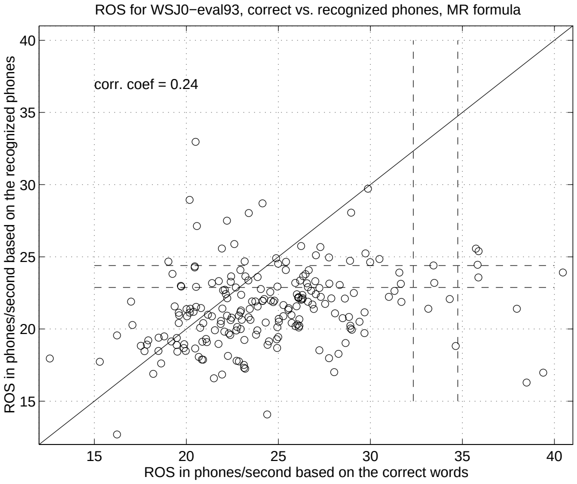


Figure 20: The plot shows the relationship between the correct word transcription method with the hypothesized phone transcription for the WSJ0-93 Eval sentences, based on the MR formula. The dashed lines are drawn at $\mu + 1.00\sigma$ and $\mu + 1.65\sigma$.

9 Conclusions

We have studied various methods of measuring ROS for the purposes of ASR. We discussed the merits of various ways of calculating the ROS for a sentence without phonetic hand transcription. We concluded that in the absence of phonetic hand transcription, using the correct word transcriptions was the best method for calculating ROS, followed by the hypothesized word and phone transcriptions. If the word recognition accuracy is acceptable, the ROS calculation based on the hypothesized word method is superior to hypothesized phone method; otherwise, the latter may be better than the former.

We also conducted a number of exploratory experiments to determine the likely sources of speech recognition errors due to unusually fast speech. We believe that the spectral features of fast and slow sounds are different, since we have been able to train classifiers to discriminate the two classes with a high degree ($\geq$ 85% for some vowels) of accuracy. This spectral difference does seem to cause higher phonetic probability estimation error rates. We also have observed an association between inappropriate word models for fast speech (due to exceptionally short phone duration or deletion) and recognition error rate.

We also implemented modifications to our ASR system to make it more robust to fast speech. We adapted our MLP phonetic probability estimator and changed the word models in our lexicon to better model the durations of fast speech. The modification with the most performance gain was obtained by modifying transitional probabilities, where the exit probabilities for the vowels were increased to 0.9, the stops to 0.7, and the rest of the phones gradually between 0.7 and 0.9. Assuming an ideal ROS estimator (which knows about the correct word transcription), the relative improvements for both fast and all sentences were significant, with $p < 0.01$ and $p < 0.05$ respectively. The relative improvement on the fast sentences were also significant ($p < 0.01$) when ROS was estimated based on the hypothesized words and phones method. The hypothesized words criterion was slightly better than hypothesized phones criterion in estimating the ROS of a novel sentence.

10 Future Directions

We suggest the following areas for future work:

- Broad category phone recognition may provide a less expensive (in terms of time and resources) alternative than phone recognition for ROS estimation. It is our guess that the accuracy of ROS estimates using the former method would be very similar to ROS calculated using the latter, since we have observed that the reduced errors in broad phonetic category recognition are mostly due to a reduced substitution rate, which does not affect the ROS measure.

- For applications where ROS must be measured in a smaller granularity than of a sentence, ROS may be measured per phone, per 1 second intervals, or per group of syllables. Distributions of this variable may be sufficient, or perhaps phone-specific measures may be required. For instance, the duration of a phone in a given utterance may be compared to the average (perhaps the context dependent average) duration of a phone, and a standardized Z value may be calculated to determine how the phone duration compares to the *ideal* phone. Since phone recognition is more error prone than broad category phone class recognition, the latter may be performed on the novel utterance instead. To get a smoothed estimate of the ROS variations along the whole utterance, the ROS may be calculated successively for overlapping time windows.

- Although rule-based pronunciation modeling did not reduce overall word error, we observed strong effects in the detailed error analysis. It may be that the improvements are cancelled when applied indiscriminately to all words. this avenue of research still seems like a likely source of improvements for conversational speech.
- Adapting the acoustic models and the word model durations improved the error for fast sentences. Combining the two methods, though, was not always beneficial. Studying the interaction between these two adaptations may lead to better robustness techniques. In particular, we are considering the use of a discriminant HMM approach [2] to simultaneously learn the acoustic and phonetic dependencies on rate.
- Even though using hypothesized phone transcriptions is a slightly inferior method to using hypothesized word transcriptions when the word recognition accuracy is acceptable, the former may be used as a faster, less expensive alternative. It is also possible to develop a general ROS detector for any given speech data, for demo purposes, for example.

As a final note, although some of the improvements may seem insignificant with respect to a large collection of sentences, an ROS-tuned system increases robustness to fast speakers, for whom the system might fail seriously. For example, for the fastest sentence in WSJ0-93 evaluation set, our baseline system has a word error of 40%. The ROS-tuned system, however, reduces this error to 20%, effectively reducing the word errors by 50%. This reduced degradation for the extreme cases could help user acceptance of ASR technology.

Acknowledgments

Thanks to the past and present members of the speech group, including Krste Asanović, Hervé Bourlard, Steve Greenberg, Dan Jurafsky, Brian Kingsbury, Phil Kohn, Yochai Konig, Florian Schiel, Gary Tajchman, and Su-Lin Wu for their insightful comments and very many discussions on this topic, to Jerry Feldman for his helpful comments on this draft, and to Matt Seigler for discussing their results. Thanks to ICSI for general support, not to mention the bi-weekly tea hour which provide a sugar, as well as a morale boost. This work was supported by NSF grant MIP-9311980 and SRI subcontract from ARPA contract MDA904-90-C-5253.

11 Appendix

Phones in the TIMIT Database					
TIMIT	IPA	Example	TIMIT	IPA	Example
pcl	p^o	(p closure)	bcl	b^o	(b closure)
tcl	t^o	(t closure)	dcl	d^o	(d closure)
kcl	k^o	(k closure)	gcl	g^o	(g closure)
p	p	**p**ea	b	b	**b**ee
t	t	**t**ea	d	d	**d**ay
k	k	**k**ey	g	g	**g**ay
q	ʔ	ba**t**	dx	ɾ	dir**t**y
ch	tʃ	**ch**oke	jh	dz	**j**oke
f	f	**f**ish	v	v	**v**ote
th	θ	**th**in	dh	ð	**th**en
s	s	**s**ound	z	z	**z**oo
sh	ʃ	**sh**out	zh	ʒ	a**z**ure
m	m	**m**oon	n	n	**n**oon
em	m̩	botto**m**	en	n̩	butto**n**
ng	ŋ	si**ng**	eng	ŋ̍	Wash**ing**ton
nx	ɾ̃	wi**nn**er	el	l̩	bott**le**
l	l	**l**ike	r	r	**r**ight
w	w	**w**ire	y	y	**y**es
hh	h	**h**ay	hv	ɦ	a**h**ead
er	ɝ	b**ir**d	axr	ɚ	butt**er**
iy	i	b**ee**t	ih	ɪ	b**i**t
ey	e	b**ai**t	eh	ɛ	b**e**t
ae	æ	b**a**t	aa	ɑ	f**a**ther
ao	ɔ	b**ou**ght	ah	ʌ	b**u**t
ow	o	b**oa**t	uh	ʊ	b**oo**k
uw	u	b**oo**t	ux	ü	t**oo**t
aw	ɑ^w	ab**ou**t	ay	ɑ^y	b**i**te
oy	ɔ^y	b**o**y	ax-h	ə̥	s**u**spect
ax	ə	**a**bout	ix	ɨ	deb**i**t
epi		(epen. sil.)	pau		(pause)
h#		(silence)			

Table 18: Phone Types Used

References

[1] Abrash, V., Franco, H., Sankar, A., Cohen, M. Connectionist Speaker Normalization and Adaptation, *The Proceedings of EUROSPEECH95*, pp. 2183–2186, Madrid, Spain, September 1995.

[2] Bourlard, H., & Morgan, N. *Connectionist Speech Recognition,* Kluwer Academic Press, 1994.

[3] Crystal, T.H., and House A.S. Segmental Durations in Connected Speech Signals: Preliminary Results. *The Journal of the Acoustical Society of America*, Vol 72, No 3, pp. 705-716, September 1982.

[4] Fisher, W. M., et al. An Acoustic-Phonetic Database, *The Journal of the Acoustical Society of America,* Suppl. 1, Vol 81, pp. S92, Spring 1987.

[5] Goldman-Eisler, F. The Determinants of the Rate of Speech Output and Their Mutual Relations. *The Journal of Psychosomatic Research.* Vol 1, pp. 137-143, 1956.

[6] Grosjean, F. Temporal Variables Within and Between Languages. *in Dechert, Raupach, Towards a cross-linguistic assessment of speech production*, Lang, Frankfurt am Main, 1980.

[7] Grosjean, F., Deschamps, A. Analyse des variables temporelles du francais spontané. II. Comparison du francais oral dans la description avec l'anglais (description) et avec le francais (interview radiophonique). *Phonetica*, Vol 28, pp. 191-226, 1973.

[8] Hermansky, H. Perceptual Linear Predictive (PLP) Analysis of Speech, *Journal of the Acoustical Society of America,* Vol 87, pp. 1738-1752, 1990.

[9] Hillenbrand, J., Getty, L.A., Clark M.J., and Wheeler K. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, Vol 97, No 5, pp. 3099–3111, May 1995.

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

[11] Kaisse, E. *Connected Speech: the Interaction of Syntax and Phonology.* Academic Press, 1985.

[12] Lamel, L.F. *Personal communication.,* September & November 1995.

[13] Lamel, L.F., Kassel, R.H., and Sennef, S. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus. *Proceedings of DARPA Speech Recognition Workshop*, L.S. Baumann (ed.), pp. 100-109, February 1986.

[14] Lindblom, B. Spectrographic Study of Vowel Reduction. *Journal of the Acoustical Society of America*, Vol 35, pp. 1773-1781, 1963.

[15] Malécot, A., Johnston, R., and Kizziar, P.-A. Syllabic Rate and Utterance length in French. *Phonetica*, Vol 26, pp. 235-251, 1972.

[16] Miller, J.L., Grosjean, F., Concetta, L. Articulation Rate and Its Variability in Spontaneous Speech: A Reanalysis and Some Implications. *Phonetica*, Vol 41, pp. 215-225, 1984.

[17] Mirghafori, N., Fosler, E. and Morgan, N. Towards Robustness to Fast Speech in ASR, *Proceedings of ICASSP '96*, Atlanta, Georgia, May 1996 (to appear).

[18] Mirghafori, N., Fosler, E., and Morgan, N. Why Is ASR Harder For Fast Speech And What Can We Do About It?, *Proceedings of the IEEE Signal Processing Society ASR Workshop at Snowbird 1995,* Snowbird, Utah, December 1995 (to appear).

[19] Mirghafori, N., Fosler, E., and Morgan, N. Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes, *The Proceedings of EUROSPEECH95*, pp. 491-494, Madrid, Spain, September 1995.

[20] Neto, J.P., Almeida, L.B., Hochberg, M., Martins, C., Nunes, L., Renals, S., Robinson, T. Speaker Adaptation for Hybrid HMM-ANN Continuous Speech Recognition System, *The Proceedings of EUROSPEECH95*, pp. 2171–2174, Madrid, Spain, September 1995.

[21] National Institute of Standards and Technology. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. NIST Speech Disc CD1-1.1, 1990.

[22] Ohde R. N., & Sharf D. J. *Phonetic Analysis of Normal and Abnormal Speech,* Merrill, 1992.

[23] Ohno, S., and Fujisaki, H. A method for Quantitative Analysis of the Local Speech Rate *The Proceedings of EUROSPEECH95*, pp. 421-424, Madrid, Spain, September 1995.

[24] Pallett, D.S., Fiscus, J.G., Fisher, W.M., Garofolo, J.S., Lund, B.A., Przybocki, M.A. 1993 WSJ-CSR Benchmark Test Results, *ARPA's Spoken Language Systems Technology Workshop*, Princeton, New Jersey, March 1994.

[25] Pallett, D.S., Fiscus, J.G., and Garofolo, J.S. Resource Management Corpus: September 1992 Test Set Benchmark Test Results, *ARPA's Continuous Speech Recognition Workshop,* Stanford, California, September 1992.

[26] Paul, D.B., and Baker, J.M. The Design for the Wall Street Journal-based CSR Corpus *Proceedings of ICSLP 92*, pp. 899-902, Alberta, Canada, October 1992.

[27] Port, R.F. Linguistic Timing Factors in Combination. *Journal of the Acoustical Society of America*, Vol 69, pp. 262-274, 1981.

[28] Robinson, T., Almeida, L., Boite, J.M., Bourlard, H., Fallside, F., Hochberg, M., Kershaw, D., Kohn, P., Konig, Y., Morgan, N., Neto, J.P., Renals, S., Saerens, M., & Wooters, C. A Neural Network Based, Speaker Independent, Large Vocabulary, Continuous Speech Recognition System: The WERNICKE Project, *Proceedings of EUROSPEECH'93*, Berlin, Germany, September 1993.

[29] Seiyama, N., Nakamura, A., Imai, A., Takagi, T., and Miyasaka, E., Portable Speech Rate Conversion System, *Proceedings of EUROSPEECH '95,* pp. 1717-1720, Madrid, Spain, september 1995.

[30] Siegler, M.A., *Personal communication.*, June 1995.

[31] Siegler, M.A., and Stern, R.M., On The Effects Of Speech Rate In Large Vocabulary Speech Recognition Systems, *Proceedings of ICASSP '95*, pp. 612-615, Detroit, Michigan, May 1995.

[32] Summerfield, Q. On Articulatory Rate and Perceptual Constancy in Phonetic Perception. *Journal of Experimental Psychology and Human Performance*, Vol 7, pp. 1074-1095, 1981.

[33] Stolcke, A., and Omohundro, S., Best-first Model Merging for Hidden Markov Model Induction, TR-94-003, ICSI, Berkeley, CA, January 1994.

[34] van Leeuwen, D.A., van den Berg, L.-G., Steeneken, H.J.M., Human Benchmarks for Speakers Independent Large Vocabulary Recognition Performance. *Proceedings of EUROSPEECH '95*, pp. 1461-1764, Madrid, Spain, september 1995.

[35] Zwicky, A. Auxiliary Reduction in English. *Linguistic Inquiry* 1.323-336, 1970.

[36] Zwicky, A. Note on a Phonological Hierarchy in English. *Linguistic Change and Generative Theory,* ed. by R Stockwell & R. Macaulay. Indiana University Press, 1972.

[37] Zwicky, A. On Casual Speech. In *Eighth Regional Meeting of the Chicago Linguistic Society,* pp. 607-615, 1972.