TOWARDS ROBUSTNESS TO FAST SPEECH IN ASR

Nikki Mirghafori *Eric Fosler* *Nelson Morgan*

International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
University of California at Berkeley, EECS Department, Berkeley, CA 94720
Tel: (510) 643-9153, FAX: (510) 643-7684, Email: {nikki, fosler, morgan}@icsi.berkeley.edu

ABSTRACT

Psychoacoustic studies show that human listeners are sensitive to speaking rate variations [10]. 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 [6]. In our earlier work [5], we studied the causes of higher error and concluded that both the *acoustic-phonetic* and the *phonological* differences are sources of higher word error rates. In this work, we have 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 also 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.

1. INTRODUCTION

There are demonstrable speaking rate differences between speakers[1]. Miller et al. [4] have shown that the articulation rate varies considerably within each 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 [7]. Listeners are extremely sensitive to these variations and treat the segmentally relevant acoustic properties in a rate-dependent manner [10]. ASR systems, perhaps even more than humans, are sensitive to the rate of speech differences, as double to quadruple word recognition error rates of average speakers have been observed for fast speakers [6].

In our earlier work [5], we investigated the *acoustic-phonetic* and the *phonological* differences of fast speech as the source of high word error, and implemented mechanisms to make our ASR system more robust to fast speech given *a priori*[2] information that enabled us to calculate the ROS. In this paper, we discuss ways to measure the ROS of a novel sentence reliably and use this information in the recognition process.

In our experiments, we use ICSI's hybrid HMM/MLP speech recognition system (explained in [1]). Since similar rate of speech effects have been observed for recognizers incorporating mixtures of Gaussians [6, 8], we think it likely that the conclusions of our work will be useful in those systems as well.

2. MEASURING THE RATE OF SPEECH

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 such a measure. It has been our experience that choosing one ROS metric over others can lead to significant differences in experimental results.

In the next two sections, we will discuss the various dimensions along which ROS can be measured and report our study on the effects of these variables.

2.1. Issues in Measuring the Rate of Speech

In this section we briefly discuss some choices that must be made in choosing a ROS measure.

2.1.1. Treatment of Mid-Sentence Silences

Should mid-sentence silences be included in the ROS calculation or dropped? For measuring ROS, we think it is of more value to exclude mid-sentence silence periods, since these durations may be dependent on factors other than speech rate. We will revisit this issue in our experiments in Section 2.2.

2.1.2. Granularity of Calculating ROS

Should ROS be measured per speaker or per sentence? The advantage of the former is that it allows for the grouping of speakers into "fast" and "slow" speakers. Speaker categorization is more intuitive, as we tend to think of speakers, and not just a set of particular sentences, as belonging to either a fast or slow group. The disadvantage is that for a given speaker, the ROS varies considerably *across* sentences [4].

2.1.3. Units of ROS

Does words/second more accurately characterize ROS or phones/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 the unit, these two sentences, which have nearly identical phonetic structure, spoken at the same speaking rate, will be labeled with widely varying ROSs.

2.1.4. Formula for Calculating the ROS

There are (at least) two ways to calculate the ROS 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 sentence [5] as in $ROS_{IMD} = \frac{n}{\sum_i duration_i}$, 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 the phone rates are averaged to get

[1] The doubtful may attend a public auction.

[2] In particular, we assumed the knowledge of the correct word transcription.

the ROS for the sentence [8], that is, $ROS_{MR} = \frac{\sum_i rate_i}{n}$, where $rate_i$ is defined as $\frac{1}{duration_i}$ for each phone.

2.1.5. Using ASR technology to estimate the ROS

How do we estimate the ROS of sentences for which we do not have the correct phone level transcription? If we do have the correct word level transcription, we can use it to perform a forced alignment to obtain the phone duration information. If we have neither the correct word or phone level transcription, there are (at least) two possible options. One is to perform *word* recognition on the novel utterance and use the hypothesized transcription for the forced alignment (also suggested by [8]). 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 wrong phonetic segmentation. A second option is to perform *phone* recognition for the novel utterance and use the state transition information to determine ROS. The advantage of this method is that we can estimate the ROS for any novel utterance, even if 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.

2.2. Correlation of the ROS Measures

We used TIMIT for the following experiments. 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. The relevant values are shown in Table 1.

As we see in Table 1, the ROSs measured using the MR formula are consistently less correlated with the phonetically hand transcribed ROS calculated using the same formula. The IMD formula seems to be a more reliable way of estimating the ROS of a sentence. Also, taking out mid-sentence silences seems to make the ROS estimation slightly more consistent. The correct word transcription method is superior to using the hypothesized phone transcriptions. 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, and the word recognition error rate is high given our simple back-off bigram grammar (estimated from the TIMIT training set). We will revisit these methods for WSJ0 in Section 4.1.

3. ANALYSIS OF FAST SPEECH

In our earlier work [5], we explored two sources for the higher error rate of faster sentences: (*acoustic-phonetic* and *phonological* causes. First, we were able to train artificial neural networks to discriminate between fast and slow frames (using PLP [2] and energy features) for a given phone and gender. The discrimination accuracy on average was about 70%, and for some vowels between 80-90%. We concluded that because of increased coarticulation effects, the spectral features of fast speech seem to be inherently different from normal speech and these differences are reflected in the extracted features (*acoustic-phonetic* causes). We also observed a strong correlation between ROS and duration and deletion mismatches (ρ = 0.93) in the word models. Therefore, the second culprit of the higher word error rates may be that the normal word models are unsuitable for fast speech because of phonemic durational mismatches (*durational errors*) or phone omission (*deletion errors*).

4. INCREASING ROBUSTNESS TO ROS

In the following two sections, we discuss our experiments in increasing robustness to fast speech. All the experiments were run on the WSJ0 corpus, and we have used the WSJ0-93 evaluation set for testing because two of the ten speakers in this test set speak very quickly and 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. The overall structure of our ROS robust system is shown in Figure 1.

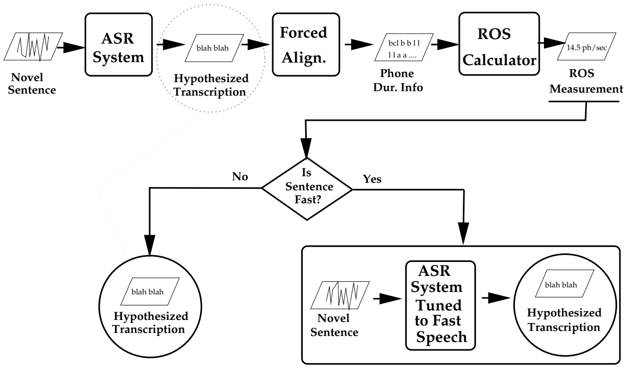


Figure 1. The over-all structure of our rapid-speech-tuned ASR system. The conditional branch is chosen on the basis of a threshold in ROS estimate.

4.1. The ROS Estimator for WSJ0

Here, we briefly look at how each of the ROS estimation methods choose a set of "fast" sentences from the WSJ0-93 evaluation set.

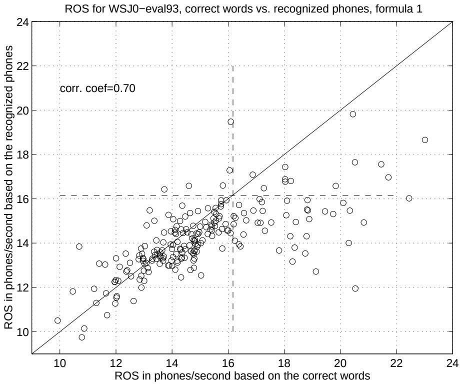


Figure 2. The correlation of the correct word transcription method with the hypothesized phone transcription for the WSJ0-93 Eval sentences, based on the IMD formula. The dashed lines are drawn at $\mu + 1.65 * \sigma$.

We see in Figures 2 and 3 that the ROS calculated using the hypothesized word transcriptions has higher correlation with the ROS calculated using the correct word transcriptions than the hypothesized phone transcriptions. As we commented earlier, for a tasks which the word recognition accuracy is acceptable, hypothesized words may provide a

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 1. Correlation coefficient for the 1344 TIMIT test sentences between various methods of calculating the ROS with the phonetically hand transcribed calculated ROS.

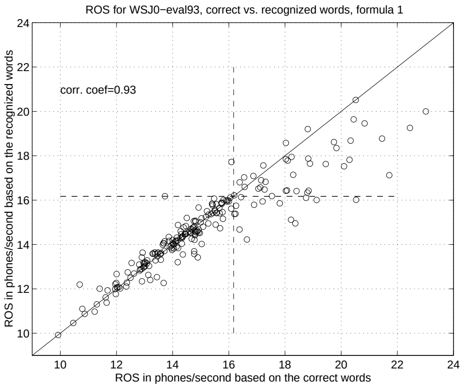


Figure 3. The correlation of the correct word transcription method with the hypothesized word transcription for the WSJ0-93 Eval sentences, based on the IMD formula. The dashed lines are drawn at $\mu + 1.65 * \sigma$.

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 the phone boundaries.

4.2. Adapting the ASR System to Fast Speech

Based on our observations in Section 3, we decided to adapt our MLP phonetic estimator to fast speech and to modify the duration constraints in the word models.

4.3. Adapting the MLP

We chose the top 5% fastest sentences (a total of 367) from the WSJ0 training data (C = ROS Cutoff = $\mu + 1.65\sigma$ = 16.17 phones/sec). We adapted our 4000 hidden unit MLP phonetic probability estimator, which was already trained on all of WSJ0, by retraining it on these fast sentences for three more epochs.

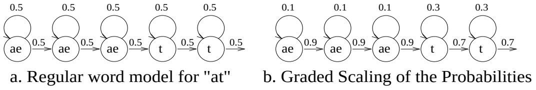
We examined the word recognition error rate on the WSJ0-93 evaluation set for the fast sentences (with $ROS > C$) and slow and medium sentences (with $ROS < C$), where C, the cutoff, was either defined to be $\mu + 1.00\sigma$ or $\mu + 1.65\sigma$ (Table 2).

From Table 2 we conclude that by lowering the ROS cutoff from 1.65σ to 1.00σ and allowing more sentences to benefit from the fast-speech modification, the overall improvement for the test set increases. Another observation from the Table 2 is that estimating the ROS using the correct word transcriptions 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 4.1.

4.4. Modeling the Duration of Fast Speech

We then adjusted the durational models of phones in order to compensate for the fast speech effects. Our current phone model, shown in Figure 4.a, requires a minimum duration constraint.

We experimented with various schemes such as reducing the number of states per phone and increasing the phone exit probability based on manners of articulation. Increasing the phone exit probability improved the error the most. The intuition behind this is that certain manners of articulation (e.g. vowels) are more likely to shorten in fast speech than others (e.g. stops) [3]. Therefore we increased the exit probabilities in a graded scale with stops at the bottom of the scale, vowels on top, and all other phones in between. The assigned probabilities for the 0.7-0.9 lexicon are reported in table 3. The scaling factor is a subjective measure of relative duration change for a particular manner of articulation. Although the scale factors have not been optimized, this scaling method improves the error rate on fast speech. The results are reported in Table 4.



a. Regular word model for "at"

b. Graded Scaling of the Probabilities

Figure 4. Examples of word models for "at".

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 3. 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].

Rel. Percent Improvement in W.E.R. for WSJ-93 Eval Set		
ROS Estimation	fast	overall (215 sents)
Corr. Word (ideal)	24.5% (50 sents)	10.6%
Hyp. Word	22.6% (37 sents)	6.2%
Hyp. Phone	23.2% (44 sents)	5.6%

Table 4. The table shows the percent improvement in recognition word error for WSJ-93 Evaluation set.

4.5. Merging the Two Solutions

We combined the most promising of the approaches we tested. We used the phonetic probabilities from the adapted net and, for decoding, used the lexicon with exit probabilities increased between 0.7 and 0.9. The improvements on this combined system was slightly less than the improvements with the tuned lexicon alone. Perhaps both modifications are making up for the same fast speech differences, and when combined together, may do "over-modification".

5. CONCLUSIONS

In earlier work [5], we conducted a number of exploratory experiments to determine the likely sources of speech recognition errors due to fast speech. We concluded that both the *acoustic-phonetic* and the *phonological* differences are sources of higher word error rates.

Relative Percent Improvement in Word Error for WSJ-93 Eval Set Using MLP Adaptation				
	C = $\mu + 1.00\sigma$		C = $\mu + 1.65\sigma$	
ROS Estimation Criteria	fast	overall (215 sents)	fast	overall (215 sents)
Correct Word (idealized)	15.0% (50 sents)	6.8%	16.7% (33 sents)	5.6%
Hypothesized Word	14.4% (37 sents)	4.3%	10.2% (21 fast)	1.9%
Hypothesized Phone	10.9% (44 sents)	3.1%	15.5% (17 fast)	1.2%

Table 2. The table shows the percent improvement in recognition word error for the WSJ-93 Evaluation set. Each row shows a different method for estimating the ROS (see text for explanation). 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.

We studied various methods of measuring ROS for a novel sentence and discussed the merits of each. We concluded that in the absence of phonetic hand transcription, using the correct word transcriptions was the best method for calculating ROS, followed by both 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. Hypothesized phone method is useful for measuring the ROS of sentences for which we do not have word models.

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.

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.

6. FUTURE DIRECTIONS

Here, we have reported a significant improvement in recognition accuracy for fast sentences. However, error rates for fast sentences are still significantly higher than for normal sentences. The following is a sketch of further research:

- 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 word error, this avenue of research still seems like a likely source of improvements for conversational speech. More specific applications of the reduced pronunciations may be required.
- 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 [1] to simultaneously learn the acoustic and phonetic dependencies on rate.

ACKNOWLEDGMENTS

Thanks to the members of the speech group, including Steve Greenberg, Dan Jurafsky Brian Kingsbury, Yochai Konig, and Su-Lin Wu for their insightful comments, and to ICSI for general support. This work was supported by NSF grant MIP-9311980 and SRI subcontract from ARPA contract MDA904-90-C-5253.

REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] Siegler, M.A.,*Personal Communication*, June 1995.

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