

ON THE EFFECTS OF SPEECH RATE IN LARGE VOCABULARY SPEECH RECOGNITION SYSTEMS

Matthew A. Siegler and Richard M. Stern
Department of Electrical and Computer Engineering
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

ABSTRACT

It is well known that a higher-than-normal speech rate will cause the rate of recognition errors in large vocabulary automatic speech recognition (ASR) systems to increase [1]. In this paper we attempt to identify and correct for errors due to fast speech. We first suggest that phone rate is a more meaningful measure of speech rate than the more common word rate. We find that when data sets are clustered according to the phone rate metric, recognition errors increase when the phone rate is more than 1 standard deviation greater than the mean. We propose three methods to improve the recognition accuracy of fast speech, each addressing different aspects of performance degradation. The first method is an implementation of Baum-Welch codebook adaptation. The second method is based on the adaptation of HMM state-transition probabilities. In the third method, the pronunciation dictionaries are modified using rule-based techniques and compound words are added. We compare improvements in recognition accuracy for each method using data sets clustered according to the phone rate metric. Adaptation of the HMM state-transition probabilities to fast speech improves recognition of fast speech by a relative amount of 4 to 6 percent.

1. INTRODUCTION

Speech rate has been shown to have a significant effect on recognition [1]. When speech rate exceeds a threshold, recognition accuracy drops. It is difficult to pinpoint this threshold as there is no standard metric for quantifying speech rate. The first part of this paper is concerned with computing speech rate and this threshold.

In addition, the compensation of fast speech can potentially make use of analyses of speech modeling errors at many levels of the recognition process, as the presence of fast speech can impair both acoustic and language models. The second part of this paper addresses a small subset of the possible compensation procedures, focussing on easily-implemented procedures that do not require complete retraining of the ASR system. Since these procedures provide only limited benefit, we also suggest a number of fruitful avenues for future research.

We used the Wall Street Journal (WSJ1) training corpus [2] which contains a total of 29,000 testing utterances. In all experiments a fast and somewhat less accurate version of the CMU SPHINX-II recognition system [3] with a 20,000-word vocabulary, 10,000

sex-dependent senone models, and a trigram language model was used.

2. MEASURES OF SPEECH RATE

Pallett *et al.* [1] used the *word rate* measure to compute the speech rate of utterances from the WSJ1 corpus. In their experiments, word rate was calculated by dividing the number words in the transcript by the total length of the utterance in minutes. We repeated these experiments and confirmed their results.

Figure 1 compares recognition errors observed for subsets of the WSJ1 corpus with different word rates. Each subset contains 100 unique utterances of the corpus having the same word rate. It can be seen that error rate increases when the word rate is greater than two standard deviations from the norm.

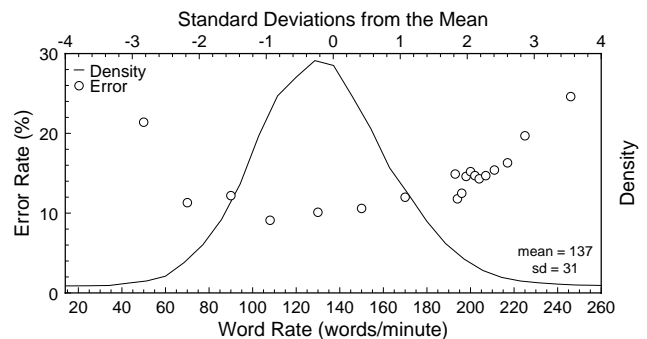


Figure 1. Recognition error rates for subsets of 100 utterances from the WSJ1 corpus grouped by word rate. The bell-shaped curve shows the distribution of utterances in the entire corpus.

It has been pointed out that word rate is unsatisfactory “because of unpredictability in the structure and length of a word, which may be monosyllabic or polysyllabic, and because of the indeterminacy of any pause durations between words” [4]. A more precise measure of speech rate must characterize the rate of information using a much smaller unit than the word. In this work we have chosen the phone as the unit of measurement.

For each phone in an utterance we define the instantaneous phone rate to be the inverse of the phone duration. In this paper the mean phone rate over the entire utterance, excluding silence periods, was used to classify an utterance. Since the phone rate is not directly related to word length, it is unaffected by the preponderance of long versus short words in a given utterance. A scatter plot of word rate versus phone rate is shown in Figure 2. The correlation between the two is rather low (0.5), so the grouping of the

corpus differs according to which measure of speech rate was used.

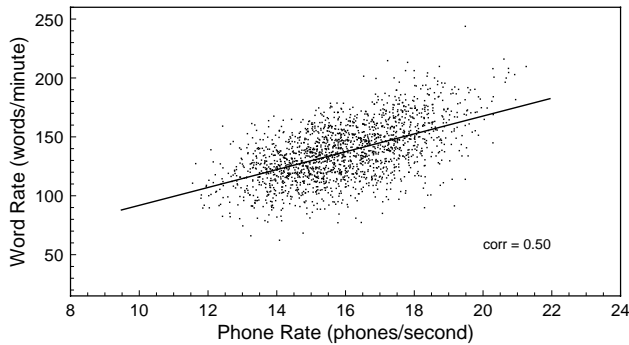


Figure 2. Scatter plot of word rate versus phone rate for utterances from the WSJ1 corpus. The line shows the best linear fit to the points.

Figure 3 shows recognition error rate as a function of phone rate, again based on subsets of 100 utterances of similar phone rate from the original WSJ1 corpus. In fast speech, increases in both deletion and substitution errors were responsible for the overall increase in error rate. In comparing Figures 1 and 3, it appears that error rate is more sensitive to phone rate than word rate. The phone rate need only be one standard deviation above the mean for error rate to increase, while the word rate must be above two standard deviations.

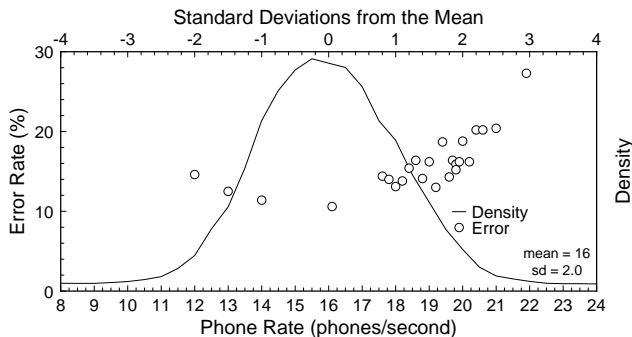


Figure 3. Recognition error rates for subsets of 100 utterances from the WSJ1 corpus grouped by phone rate. The bell-shaped curve shows the distribution of utterances in the entire corpus. Please note that the data sets are different than in Figure 1.

Computation of both the word rate and phone rate requires the correct transcription for the utterance. To compute phone rate, forced alignment is first used to determine the phone segmentation and then phone durations are computed from this segmentation.

In many situations, it is important to be able to detect fast speech without knowing the transcript *a priori*. We investigated the use of the recognition system’s transcript hypothesis containing time-alignment information. Figure 4 compares estimated phone rates (produced by the recognition system) to actual phone rates (derived from the written transcripts). Although negatively biased, the estimate of the phone rate is monotonically related to the actual phone rate, and therefore can be used to predict the presence of fast speech.

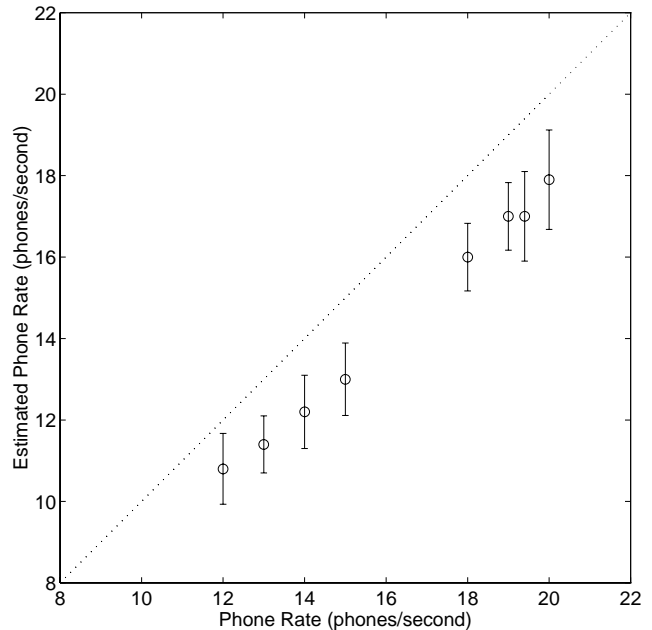


Figure 4. Estimated versus actual phone rates for each of eight 100-utterance sets having the same actual phone rate. Error bars represent ± 1 standard deviation in the estimate. The dotted line shows the unbiased predictor.

3. COMPENSATION FOR THE EFFECTS OF HIGH SPEECH RATE

Degradation of recognition accuracy due to high speech rate may occur when the acoustical models and language models obtained from training with average speech fail to describe the corresponding characteristics of fast speech. In this section, we describe modifications made to three components of the recognition system to compensate for the effects of high speech rate: the models of the acoustical characteristics of speech sounds, the models of the HMM state-transition probabilities, and the pronunciations of words in the dictionary. These procedures were selected for consideration in this work because they are relatively easy to implement without complete retraining of the speech recognition system. Since we find, unfortunately, that these methods provide only very limited benefit, we also suggest several other more computationally-intensive approaches that may be more effective.

3.1. Modification of Acoustic Models

If the production of fast speech differs from the production of speech at a normal rate, the acoustical characteristics of the output will differ as well. In previous research, recognition accuracy has been improved through the use of VQ codebooks that were specific to gender, pitch, and environment (*e.g.* [5,8]). We developed rate-specific codebooks by performing Baum-Welch codebook reestimation for fast speech. Figure 5 is a scatter plot comparing the values obtained for the C-0 and C-1 VQ codeword locations, for the original speech and for speech derived from the fastest 1000 utterances in the database. As can be seen, there is little difference between the two sets of codeword locations. The higher-order coefficients of the two codebooks were even more similar.

Recognition accuracy of fast speech using the codebooks derived from fast speech did not improve compared to baseline accuracy.

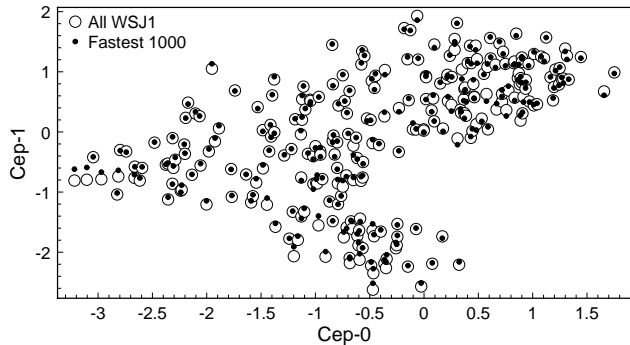


Figure 5. Codebooks entries for baseline, and fastest 1000 utterances computed from Baum Welch adaptation.

The fact that the use of rate-specific codebooks was unsuccessful in this experiment suggests that the long-term average acoustical characteristics of normal speech and fast speech are similar. However, it is also known that codebook variations can depend on phonetic class [e.g. 6], so it is possible that the use of codebook modifications based on phoneme class as well as on speech rate may be more successful.

3.2. Modification of HMM State-Transition Probabilities

Since faster speech is typically less carefully articulated, it is expected that recognition accuracy could be improved by modifying representations of duration of the various phonetic productions in the HMMs. It has been shown that when speech rate increases the change in duration of vowels is greatest [7]. We used forced-alignment techniques to confirm this observation for the WSJ1 corpus. Obtaining state and phone segmentations for all utterances, we noticed a very high occurrence of extremely short (30-ms) durations in the fast speech.

Figure 6 compares histograms of vowel durations for normal and fast speech in the WSJ1 corpus, along with the duration statistics of vowel segments in the original HMM representation for normal speech. It can be seen that the phone durations of the vowel segments of the HMMs more closely resemble average than fast speech.

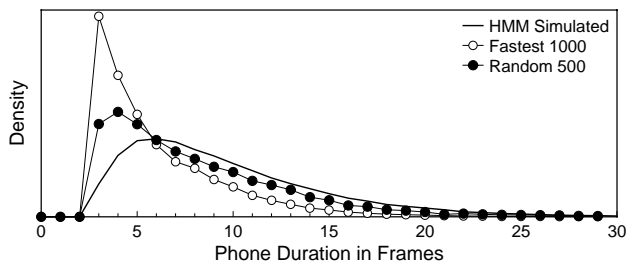


Figure 6. Histogram of vowel durations for a random set of 500 utterances, the fastest 1000, and as modeled by the HMMs.

Based on the comparisons shown in Figure 6, we believe that more explicit modeling of phone duration would improve recognition accuracy, which can be most easily accomplished by modification of the state-transition probabilities in the HMMs. In the SPHINX-II system, only monophone transition probabilities are trained because they are assumed to be relatively unimportant in recognition [8].

We evaluated the importance of state-transition probabilities by creating two new sets of state-transition models. The state-transition probabilities in the first set were made equal, which tended to shorten the average model phone duration. The second set of models had transition probabilities adapted to the fastest 1000 utterances, by counting state-transitions from segmentation of these utterances. Recognition error rates obtained using these probabilities are compared with the original baseline probabilities in Figure 7. The relative decrease in error rate obtained using rate-specific unequal probabilities is 4 percent for speech with phone rates from 18 to 22 phones/second, and 6 percent for speech with phone rates from 20 to 22 phones/second. Equal transition probabilities provided a lower error rate for fast speech, but they were of no impact on normal-rate speech. The adapted models provide an even lower error rate for fast speech, and a slight increase in error rate for speech spoken at the normal rate.

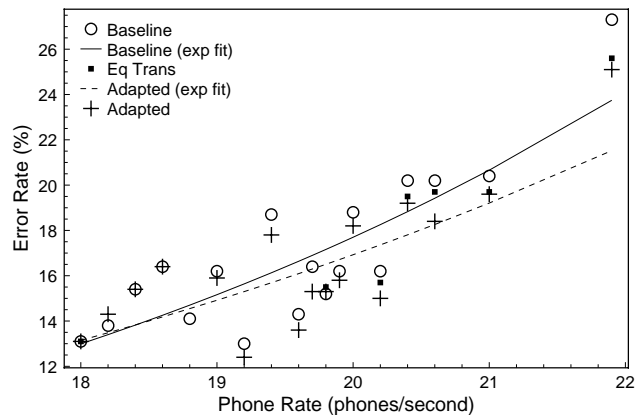


Figure 7. Comparison of recognition error rates using the original state-transition probabilities, equal transition probabilities, and probabilities adapted to fast speech. Smooth curves indicate the best exponential fit to the baseline and adapted models.

3.3. Modification of Pronunciation Dictionaries

Fast speech frequently produces changes in word pronunciation as well as in phone articulation, and it has been shown that these changes occur both within and between words [9]. In this section we focus on both intra-word and inter-word transformations in fast speech, and methods to compensate for them.

3.3.1 Intra-Word Transformations

A speech recognition system developed using average-rate speech may not be able to identify a rapidly-spoken word spoken if the dictionary entry for that word is different than as pronounced. A linguist helped us to identify possible pronunciation differences for fast speech. The rules governing intra-word transformations are numerous and complex [9].

Because we found deletions of unstressed vowels to be very common in fast speech, we selected two simple rules involving transformations of the *schwa*. In each case, we modified the dictionary and repeated recognition of the fast speech. The rules were:

1. Eliminate the *schwa* between two consonants.
2. Eliminate all non-initial and non-final *schwas*.

300 of the 20,000 words in the database were modified by Rule 1, and 7,000 words were modified by Rule 2. For each rule, recogni-

tion was performed using the new dictionary, and a union of the new and original dictionary. Neither rule resulted in a significant change in the overall word error rate.

3.3.2 Inter-Word Transformations

The pronunciation of a sequence of words is often different than the words spoken in isolation [8], and if they are of short duration and have few phones, deletions can occur. In our observations of recognition performance of fast speech, the fraction of errors that are deletions is approximately 33 percent. In comparison, word deletions comprise only 5 to 10 percent of all errors for normal speech. For the fast-speech utterances that we examined, only 10 different spoken words, *THE*, *AND*, *TO*, *A*, *OF*, *IN*, *THAT*, *WERE*, *ARE*, and *I*, represented 55 percent of all word deletions even though they represented only 20 percent of all words in the transcripts. 33 percent of these deletions were merges of the form

“*X Y*” -> “*X*” and “*X Y*” -> “*Z*”

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

“*OF THE*” -> “*OF*” and “*AND A*” -> “*THE*”

Because it is not possible to reduce deletion errors due to merges of short words by changing their individual pronunciation, we added compound words to the dictionary instead. These compound words are of the form “*IN_THE*”, “*AND_IN*” and have slightly different pronunciations than each word separately. From 20,000 words in the dictionary, 164 new compounds representing the most frequent merges were added.

Other sites [10] have obtained a slight decrease in error rate using dictionaries with compound words for the Switchboard corpus [11]. Nevertheless, we found that adding compound words to our dictionary did not improve recognition accuracy.

4. DISCUSSION

In this paper, we selected phone rate over word rate as a more precise measure of speech rate and found that recognition errors increase when the phone rate exceeds 1 standard deviation from the mean. As some studies have found that vowel durations are most sensitive to speech rate [7], it is also possible the average vowel rate would be a superior metric.

We explored several methods in three different domains of speech modelling to reduce recognition errors for fast speech. Although the first method, codebook adaptation, failed to improve the recognition performance of fast speech, we believe better results may be obtained using a combination of separate phone-dependent and rate-dependent codebooks.

The second method, HMM state-transition probability adaptation, demonstrated that state-transition probabilities do indeed affect recognition, and that adaptation can reduce error rates for fast speech by a relative amount of 4 to 6 percent.

We applied a small number of rules that manipulated a large number of pronunciations to compensate for some of the effects of intra-word transformations. While these simple rules did not yield improvements in accuracy, it is possible that such an approach could be more successful with a more complete and systematic set of transformation rules.

Finally, we added compound words to the dictionary based on observations of pronunciations of pairs of short words to compensate for inter-word transformations. While we observed no increase in recognition accuracy, retraining the acoustic models after manipulating the dictionary should reduce recognition errors.

5. ACKNOWLEDGEMENTS

This research was sponsored by the Advanced Research Projects Agency (DOD) under contract N00039-91-C-0158. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Advanced Research Projects Agency, or the US Government. We thank Bob Weide for his help and advice in developing rule-based modifications of the pronunciation dictionaries. We also thank Uday Jain, Pedro Moreno, Bhiksha Raj, Raj Reddy, and the rest of the speech group for numerous other contributions to this work.

REFERENCES

1. D. S. Pallett, et. al., “Be Sure to Read the Fine Print: II”, *Proc. ARPA Spoken Language Systems and Technology Workshop*, March 1994.
2. D. Paul, and J. Baker, “The Design of the Wall Street Journal-based CSR Corpus”, *Proc. DARPA Speech and Natural Language Workshop*, Feb. 1992.
3. X. Huang, F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, R. Rosenfeld, “The SPHINX-II Speech Recognition System: An Overview”, *Computer Speech and Language*, **7**: 137-48, 1993.
4. W. N. Campbell, “Extracting Speech-Rate Values from a Real-Speech Database”, *Proc. ICASSP-88*, April 1988.
5. F.-H. Liu, *Environmental Adaptation for Robust Speech Recognition*, Ph.D. Thesis, Department of Electrical and Computer Engineering, Carnegie Mellon University, June 1994.
6. F. Liu, P. Moreno, R. Stern, and A. Acero, “Signal Processing for Robust Speech Recognition”, *Proc. ARPA Human Language Technology Workshop*, March 1994.
7. G. Peterson, I. Lehiste, “Duration of syllable nuclei in English,” *J. Acoust. Soc. Am.* **32**: 693-703, 1960.
8. M.-H. Hwang, *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Dec. 1993.
9. L. R. Shockey, *Phonetic and Phonological Properties of Connected Speech*, Ph.D. Thesis, Phonetics and Phonology, Ohio State University, 1973.
10. Personal communication with Ellen Eide, BBN, Aug. 1994.
11. J.J. Godfrey, E. C. Holliman, J. McDaniel, “SWITCHBOARD: Telephone Speech Corpus for Research and Development”, *Proc. ICASSP-92*, March 1992.