Presented at Eurospeech-97, Rhodes, vol. 4 pp. 2079-2082.

SPEECH RECOGNITION USING ON-LINE ESTIMATION OF SPEAKING RATE

Nelson Morgan, Eric Fosler, and Nikki Mirghafori

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: {morgan, fosler, nikki}@icsi.berkeley.edu

ABSTRACT

In this paper, we describe a rate of speech estimator that is derived directly from the acoustic signal. This measure has been developed as an alternative to lexical measures of speaking rate such as phones or syllables per second, which, in previous work, we estimated using a first recognition pass; the accuracy of our earlier lexical rate estimate depended on the quality of recognition. Here we show that our new measure is a good predictor of word error rate, and in addition, correlates moderately well with lexical speech rate. We also show that a simple modification of the model transition probabilities based on this measure can reduce the error rate almost as much as using lexical phones per second calculated from manually transcribed data. When we categorized test utterances based on speaking rate thresholds computed from the training set, we observed that a different transition probability value was required to minimize the error rate in each speaking rate bin. However, the reduction of error provided by this approach is still small in comparison with the increases in error observed for unusually fast or slow speech.

1. INTRODUCTION

In previous work we [1] and others [3, 4] have observed strong correlations between the performance of speech recognition systems and deviation of the test data from the average rate of speech observed in training data. We have also documented that some fairly simple changes to the system (e.g., modification of state transition probabilities) can be used to improve performance in this case. However, our strategy was based on a *lexically-based* measure of speaking rate. The best measures that we found were based on counting phonetic units following recognition. This required a recognizer that performed reasonably well in the first place. For difficult problems such as the recognition of conversational speech, performance may be too poor for this to be a reliable technique. One solution to this problem is to try to estimate phone boundaries without using a full recognizer; for instance, Verhasselt and Martens [5] employed a multi-layer perceptron approach with measurable success on the TIMIT database.

However, in conversational speech common phonetic units corresponding to allophones or allophone segments may be significantly transformed or even disappear, and so may not be reliable measures of speaking rate. We believe that it would be desirable to develop a measure of speaking rate that is directly based on signal processing of the speech, without reference to lexical units or requiring the use of a recognition system for estimation. Such a measure should also be continuously computable, and yet correlate reasonably well with lexically-based measures such as phone count (although some errors are inevitable due to the phone transformation mentioned above).

In the work described here, we have begun to address these concerns. In particular, we have developed a simple estimator of speech rate based on 1-2 seconds of the speech signal. We have computed the measure for a sample of data from the Switchboard Corpus that has been phonetically transcribed, as well as for a phonetically transcribed portion of the OGI Numbers corpus from the Oregon Graduate Institute. In each case we have determined the correlation coefficient between a windowed phone count and the new measure for this sample. We also have examined the relationship between this measure and word error rate. Finally, we have implemented a preliminary recognition strategy in which we only modify the transition probabilities of a hybrid system given the rate estimate, and have observed results for the Numbers corpus.

2. SIGNAL ANALYSIS

Previous work [2] has demonstrated that the speech signal is significantly altered for varying speaking rates. The most obvious change to the speech, however, is the variation in the energy envelope. In other words, the energy envelope of the speech simply has more rapid change when the speaking rate is higher. This should be reflected in the short term spectrum of the energy envelope. Finer (more frequency-dependent) measures could potentially provide higher accuracy, but as a first attempt the wideband measure should incorporate the gross properties of speaking rate.

We have experimented with such a measure, which we currently refer to as the energy rate or **enrate**. Currently, our basic analysis steps are:

1. Half-wave rectify the signal waveform.
2. Low-pass filter the rectified waveform (currently with a single real pole at 16 Hz).
3. Downsample to 100 Hz.
4. Hamming window 1-2 seconds of speech (one second windows provide better dynamics, 2 second windows are more stable). Step the windows with significant overlap (e.g., >75%).
5. Compute a short-term spectrum. Currently this is done using a DFT (we only use values up to 16 Hz, so a full FFT is not done).
6. Compute a spectral moment (index-weighting each power spectral value and summing), ignoring the first few spectral values (i.e., ignoring d.c.).

During development we observed that the enrate's behavior was roughly comparable to a syllabic rate.

It may be desirable to change one or more aspects of the enrate — for instance, using an adaptive filter to estimate a single best-fit resonance might be preferable to using the DFT. However, it currently appears that even this very simple measure matches the gross rate properties reasonably well. To demonstrate this, we have experimented with enrate on the Switchboard conversational speech corpus, comparing with manual markings done at ICSI.

3. EXPERIMENT 1: CORRELATION WITH LEXICAL SPEAKING RATE

The rate analysis was performed on 451 segmented utterances that are part of the 1996 development test partition used at the Johns Hopkins 1996 Workshop on Conversational Speech Recognition. These utterances had previously been manually annotated by Steve Greenberg and his students at Berkeley, so that phonetic alignments were available. Phonetic transitions were counted from these annotations using a two second window that was stepped every 10 msec, and this count was used as a local lexically-based measure of speaking rate. A two second window was used for

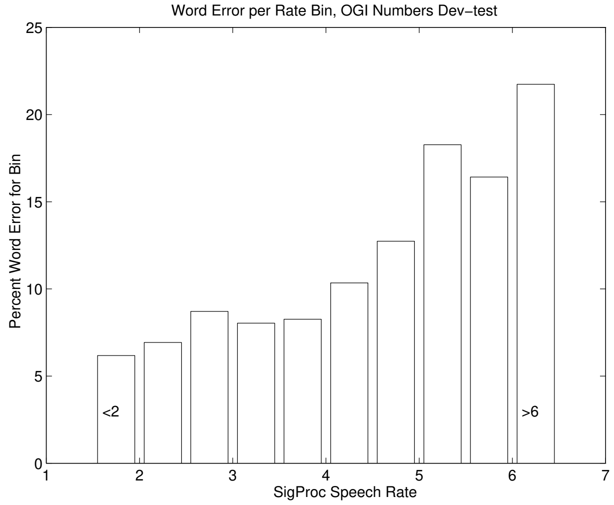

Figure 1: Increasing error rates with faster speech

rate estimation. The initial and final 100 frames (for which a full window was not available) were assigned the value of the closest valid 2 second window. This measure was then compared with the corresponding measure from the signal processing described in the previous section. Two colleagues at ICSI, Dan Ellis and Joy Hollenback, also provided a rough syllabification of the development sentences; the same lexical rate measure was applied in order to determine the syllabic rate using the same windowing criterion.

4. RESULTS AND DISCUSSION

Correlations were calculated between enrate and windowed lexical (phone and syllable) measures of speech rate; 136782 frames of conversational telephone speech from development test data in the Switchboard corpus were used. Surprisingly, the match was better between the energy-based rate measure and the phonetic rate (correlation=0.50), rather than the syllabic (correlation=0.42). However, this may be due to inaccuracies in the syllabic markings, which were obtained by a semiautomatic procedure. The phonetic markings, on the other hand, were manually generated. The correlations are strongly significant, but clearly the new measure is at present fairly noisy.

In a follow-up experiment, we computed the enrate and a corresponding phone-based measure derived from automatic alignments on utterances from the OGI Numbers corpus using a window covering the whole utterance, and got quite similar results (correlations of roughly 0.5 between the phonetic counts and new rate measure). In addition, we calculated the average error rate for different speaking rates for our baseline system (Figure 1), and found that a high speaking rate measurement was a good predictor of increased errors, even for our crude estimator.

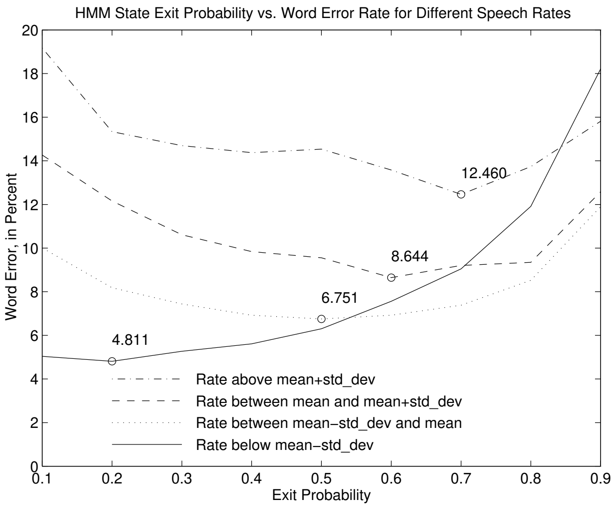

Figure 2: Response of development test set to transition probabilities. From slowest to fastest, the n in each bin were 873, 1748, 1423, and 626.

5. EXPERIMENT 2: INCORPORATING SPEAKING RATE IN RECOGNITION

As noted above, in previous work on speaking rate we found that we were able to significantly reduce the word error rate for rapidly spoken sentences using a gross measure of the average phone rate for an utterance to simply classify the utterance as fast or not. We were able to do this either by modifying the state transition probabilities in our hybrid ANN/HMM system or by altering the acoustic probabilities. In that study (which focused on sentences read from the Wall Street Journal), the transition probability modification was the most effective; for instance, increasing the probability of exiting a state reduced the errors on fast speakers significantly.

More recently, we have incorporated the enrate measure to experimentally determine the best transition probabilities for different speaking rates in the OGI Numbers corpus. We calculated the enrate for each utterance of the corpus.[1] Using the mean and standard deviation of speaking rate from the training set, the development test set was divided into four bins.

The baseline duration model of our recognizer fixes the number of states per phone to correspond to half of the mean duration of the phone, but sets the transition probabilities between states to 0.5. For recognition experiments using utterances with moderate speaking rates, this choice was roughly as good as using transition probability estimates that were more model-specific. To find an upper bound for improvements based on coarse changes to a single exit probability for all phones, we changed the exit probability for all HMMs to favor or disfavor faster speech; the word error rate for the rate-based test set bins given different state exit probability settings can be seen in Figure 2. The optimal exit transition probability (indicated by the circles on each line) increases with speaking rate, giving a 14% and 24% relative improvement for the fastest and slowest utterances respectively, in comparison to the baseline system.

How does the probability on the abscissa correspond to the average durations in the four rate bins for the training set? If we could learn a consistent relation between durations and this probability, then we could set the exit probability for each estimated speaking rate. Assume a model i with k_i left-to-right states with no skips permitted, but with a self-loop permitted on the last state only. If the optimum exit probability corresponded to matching the average duration, then

$$p_i^{exit} = \frac{1}{\mu_i - (k_i - 1)} \qquad (1)$$

This can be easily derived by taking the expected value of the exponential duration function.

As noted above, for our systems, the number of states k_i is chosen to be half of the average phone duration over the whole training set; this has been found to be a good rule-of-thumb approach to setting minimum durations for our system. To derive the exit probability automatically one could imagine separately using the duration μ_i for each bin determined from the enrate measure on the training set (as described in Figure 2).

When this relation is used, the exit probability (based on matching mean durations) varies between roughly .1 and .2 for the slowest and fastest bins respectively. This is not a good match to the exit probabilities that we observe to be the best for recognition (as can be seen from Figure 2). Further inspection of the duration distributions show that rapid speech exhibits a much more skewed characteristic than slow speech. This suggests that it would be better to incorporate explicit duration modeling for the individual bins, rather than to simply use implicit exponential models that match the mean of the sample distribution.

For the experiments illustrated in Figure 2, we would like to know how much the noisy estimation of speaking rate limits performance (since, after all, the enrate has only a .5 correlation with the phonetic rate). To examine this, we redid the experiment using manually transcribed phone sequences to derive a lexical measure of rate, split up the training set as before and adjusted the exit probability separately for each partition to minimize the error rate. The results are shown in Table 1. It can be seen that while the enrate-based procedure doesn't reduce the error as much as a lexical measure based on manual phonetic

[1] To calculate one rate measure per utterance, the signal processing window was enlarged to cover the entire utterance.

Exit probability	Word Error Rate
Optimum overall	8.57%
Enrate-derived	7.73%
Manual transcription-derived	7.54%

Table 1: Word error rates for OGI Numbers development set. Error rates are computed over complete set of 4670 words.

transcriptions, it is pretty close (and you don't need a good estimate of the phonetic transcription!) In both cases, the error rate reduction for the bin with the most rapid speech was roughly 14% in relative terms (though this figure is not directly comparable since the two rate-based procedures result in different criteria to classify an utterance as "fast".) However, preliminary results using exit probabilities set from the development set have not yet yielded much improvement on an independent test set, even for the case of rates derived from manual transcriptions. It also appears that this lack of generalization was not due to a mismatch of the speaking rate thresholds.

6. DISCUSSION

In this paper we have shown that, without estimating the number of lexical units per second directly (i.e., through recognition), we can estimate speaking rate from the acoustic signal. We have also noted that we can reduce the recognition error rate by choosing different HMM exit probabilities for utterances with different enrate values. While enrate is only moderately correlated with lexically-based measures, it appears to work well enough for this partitioning task. The exit probability was chosen as a single global parameter. It is likely that we will need more detailed durational models (e.g., altering transition probabilities separately for different broad classes) in order to generalize better. Also, we know durational effects to be accompanied by other phenomena that we have observed in rapid (or unusually slow) speech, but have not yet incorporated. In particular, the current study doesn't use any modification of the pronunciations, nor are the emission probabilities adapted as we have done in previous studies using lexical measures of rate. We are also interested in exploring the applicability of statistical models that jointly model emission and transition, since we currently don't know how to jointly compensate for these effects with the models described here.

7. ACKNOWLEDGEMENTS

This work was started in connection with the 1996 Johns Hopkins Workshop on conversational speech recognition, and was continued with the assistance of further funding from Hopkins. We also benefited from discussions about our speaking rate measure with others at the Workshop, particularly Mari Ostendorf, Liz Shriberg, and Barbara Wheatley. We also thank Hervé Bourlard for suggesting that we seek a rate measurement from the signal, and Steve Greenberg for continuing to point us in the direction of incorporating modulation-spectrum-related measures for purposes such as the one described here. We also thank our colleagues in the Switchboard Transcription Project (also led by Steve Greenberg at ICSI) for generating the manual transcriptions used in our Switchboard measurements, and Ron Cole's team at OGI for providing us with the Numbers corpus and its phonetic transcriptions. As noted in the text, others at ICSI (in particular Joy Hollenback and Dan Ellis) were also helpful in this work.

8. REFERENCES

[1] Mirghafori, N., Fosler, E., and Morgan, N., Towards Robustness to Fast Speech in ASR, *ICASSP '96*, pp. I335-338, Atlanta, Georgia, May 1996.

[2] Mirghafori, N., Fosler, E., and Morgan, N., Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes, *EUROSPEECH '95*, pp. 491-494, Madrid, Spain, September 1995.

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

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

[5] Verhasselt, J.P., and Martens, J-P., A Fast and Reliable Rate of Speech Detector, *ICSLP '96*, pp. 2258-2261, Philadelphia, Pennsylvania, October 1996.