COMBINING MULTIPLE ESTIMATORS OF SPEAKING RATE

Nelson Morgan and Eric Fosler-Lussier

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}@icsi.berkeley.edu

ABSTRACT

We report progress in the development of a measure of speaking rate that is computed from the acoustic signal. The newest form of our analysis incorporates multiple estimates of rate; besides the spectral moment for a full-band energy envelope that we have previously reported, we also used pointwise correlation between pairs of compressed sub-band energy envelopes. The complete measure, called *mrate*, has been compared to a reference syllable rate derived from a manually transcribed subset of the Switchboard database. The correlation with transcribed syllable rate is significantly higher than our earlier measure; estimates are typically within 1-2 syllables/second of the reference syllable rate. We conclude by assessing the use of *mrate* as a detector for rapid speech.

1. INTRODUCTION

We and others have been looking at the effects of speaking rate in continuous speech for several years, as reported in e.g., [6], [4]. We have found that there are strong measurable effects on acoustic distributions and on durations. We have also begun to explore effects on pronunciation; for some words, pronunciation probabilities can change significantly due to rate, e.g. for the monosyllabic word shown in Table 1. All of these effects are reflected in the word error rates of automatic speech recognition systems, which we and others have consistently shown to be significantly affected by speaking rate.

In order to explore the effects of speaking rate on recognition, we must first define a reference measure. We have found that syllables per second over a speech "spurt" (between-pause region) appears to be a reasonable measure. In the case of syllabically transcribed data, this is derived by simply counting the syllables and dividing by the segment length. As a reference point, we have used this measure for some of the experiments described in this paper. Of course, syllabic transcriptions are not accessible during the operation of a speech recognition system, so it is necessary to design a measure that can be computed without a transcription of what was said. In [4] we described the tactic of running the recognizer twice, using the first pass to hypothesize sound unit boundaries and hence the speaking rate, which would then be incorporated in a second pass; when the rate was only going to be classified as "fast" or "not fast" this could also be done by running two recognizers and taking the most probable result. However, aside from the additional computation, this method requires the assumption that the speaking rate determined by a potentially errorful recognition hypothesis would be sufficiently accurate. For difficult tasks such as conversational speech recognition, this is often not the case, particularly for unusually fast or slow speech.

pronunciation	low rate	high rate
b ih n	0.6087	0.3636
other	0.3913	0.6364

Table 1: Probabilities for canonical and non-canonical pronunciations for the monosyllabic word "been", evaluated for one hour of Switchboard speech. The threshold between "high" and "low" was the median syllable rate for the between-pause spurt containing each target syllable, and was computed from manual transcriptions.

Consequently, we have also worked to develop a measure of speaking rate that is only dependent on the acoustic signal. We previously reported in [5] an estimator called *enrate*, which was essentially the first spectral moment of the broad-band energy envelope. Kitazawa [3] also reported a full-band measure that was very similar (taking the dominant spectral peak of the long term envelope spectrum, rather than the moment). *Enrate* was correlated with transcribed syllabic rate, but the deviations were large. Since that work, we have developed an improved measure that we have called *mrate*, short for *m*ultiple *rate* estimator. *Mrate*, described below, appears to be a much closer match to transcribed syllable rate than *enrate*.

Mrate incorporates multiple estimators, a technique that has been shown to be beneficial for many aspects of speech analysis, such as the parallel pitch detection approach developed by Gold and Rabiner 30 years ago [1]. As in that case, using multiple estimators significantly improves the error variance.

2. SIGNAL PROCESSING METHODS

As noted above, we have previously used *enrate* as a measure of speaking rate [5]. *Enrate* is the first spectral moment of the wide-band energy envelope, typically computed over 1-2 seconds and restricted to roughly the spectral range from 1 to 16 Hz. That is, for $x(n)$ a half-wave rectified and low-pass filtered speech signal, $w(n)$ a Hamming window for the analysis region, and $Y(k) = DFT(w(n)x(n))$, we compute

$$enrate = \frac{\sum_{k=s}^{K} k\,|Y(k)|^2}{\sum_{k=s}^{K} |Y(k)|^2} \qquad (1)$$

Note that the result is in units of frequency increments defined by the inverse of the analysis window length — for a 1 second window, the result is in Hz. The starting analysis point s is chosen

0-7803-4428-6/98 $10.00

to reduce the effect from dc content in the energy envelope as measured by the windowed signal (typically 2 bins for a Hamming window), and endpoint K is chosen to correspond to 16 Hz.

We have found *enrate* to be useful in characterizing some of the properties of conversational speech. However, for a one hour subset of the manually transcribed Switchboard data, we found that the correlation between transcribed syllable rate and *enrate* was only about .4 (when both were measured over between-pause spurts). After a range of experiments, we have found that an average between *enrate* and two different peak-counting estimators gives us much better performance, showing a correlation of over .6 on the same data set. We are dubbing the new measure *mrate* for its use of multiple rate estimators. The second estimator used in the average is a simple peak counting algorithm performed on the wide-band energy envelope. The most effective of the three is a sub-band-based module that computes a trajectory that is the average product over all pairs of compressed sub-band energy trajectories. That is, if $x_i(n)$ is the compressed energy envelope of the i^{th} spectral band, we define a new trajectory $y(n)$ as

$$y(n) = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} x_i(n) x_j(n) \qquad (2)$$

where N is the number of bands and $M = \frac{(N)(N-1)}{2}$ is the number of unique pairs.

Peak counting is also used in this module. While this approach is already more accurate than *enrate*, averaging with the other two appeared to be beneficial. Therefore, it is this average measure that we use in the experiments described below. See Figure 1 for a block diagram of the process.

In preliminary experiments, we worked with smaller subsets of manually transcribed Switchboard data. Some of the decisions that we made based on the performance for this reduced set included:

1. Cube root compression for the sub-band energy envelopes appeared to be more effective than other compressions tried (square root or 4th root).
2. Relatively gentle modulation filters appeared to give better performance than very steep filters.
3. The average product of all sub-band energy pairs (6 combinations for 4 bands) appeared to work more reliably than summing compressed energies over the bands.
4. Normalization within each band was necessary. The resulting measure was independent of both overall energy and spectral slope.

Given these design decisions, we found that the average of *enrate* and the full-band and sub-band-based peak counting measures was significantly more correlated to transcribed syllable rate than any of the measures by themselves. Interestingly, the simple average gave essentially the same correlation as the best least squares weighting of the three component estimates, although the latter computation yielded very uneven weights.

We then applied the estimators to a much larger set of manually transcribed Switchboard syllables, as described below.

3. EXPERIMENTAL METHODS

The data used are from 5757 utterances found in the Switchboard corpus, comprising approximately four hours of data. These utterances were phonetically hand transcribed by linguists in the Switchboard Transcription Project at ICSI [2]; the transcriptions consisted of phonetic identities with syllabic boundary markings, as phonetic boundaries were often difficult to determine in this conversational speech context.

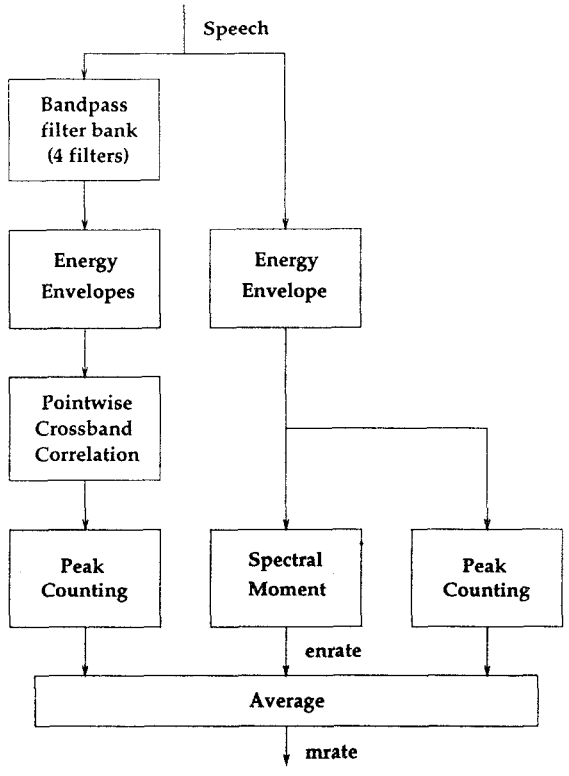

Figure 1: Major steps in the calculation of *mrate*. The band-pass processing currently uses steep FIR filters with band edges of (300,800), (800,1500), (1500,2500), and (2500,4000). Estimation is typically done over 1-2 seconds; for the experiments reported here, we used between-pause intervals, which varied from .25 to 6 seconds, but which were most typically between .5 and 2 seconds.

Utterances were segmented into spurt regions using the transcribers' pause markings, providing 7994 regions for analysis. A transcribed syllable rate was computed by dividing the number of syllables occurring in the region by the length of the spurt. We also computed *enrate* and *mrate* for each region. A separate 22-minute portion of the development test set was also prepared in the same way, providing 872 segments of test material from 441 utterances.

In order to assess the improvement of *mrate* over *enrate*, we treated the transcribed syllable rate as a gold standard, and computed correlations between each of the estimates and the syllable rate. We also determined the difference between the standard and the estimates in order to find the bias and variance of each measure.

Finally, we attempted to use *mrate* as a predictor of speaking rate in a three-class problem. The training set was divided into three equal parts (slow, medium, and fast) based on transcribed syllable rate. We then used the corresponding partition points set to determine rate classes for the development test set. In a first test, we generated a Receiver Operating Characteristic (ROC) curve for the problem of detecting "fast" speech, as defined by the test set labels for the 3 rate classes. To build this curve for a rate measure

measure	correlation	mean error	stddev error
enrate	.415	.747	1.405
sub-*mrate*	.637	.530	1.219
mrate	.671	.464	1.121

Table 2: Relationship between syllable rate computed from manual transcriptions and three signal processing measures. Results were computed for roughly 3 hours of between-pause segments taken from 4 hours of conversational speech in the Switchboard corpus.

under test, we varied its threshold and assessed correct detection percentages; non-fast segments that fell above the threshold were counted as false positives.

Since recognition systems are often adapted towards faster or slower speech (instead of building completely separate models for speech extremes), medium-rate speech that is misclassified as fast often does not have as devastating an effect as slow speech that the detector mis-classifies as fast. Therefore, we also computed ROC curves ignoring the effect of mid-speed segments to assess the tradeoffs between the accuracy of fast speech detection and the frequency of detections that are strongly in error.

4. RESULTS AND DISCUSSION

Figure 2 shows a scatter plot of *mrate* versus transcribed syllable rate for the roughly 8000 between-pause spurts. As shown in table 2, the correlation between the two measures is about .67, which roughly corresponds to 45% of the *mrate* variance being accounted for by a linear relationship with transcribed syllable rate. The overall diagonal trend is discernible in the scatter plot, though it is apparent that much of the variance is due to other factors than the transcribed rate. We have noted in a number of individual cases that a high speaking rate sometimes results in the smearing together of energy peaks, even in sub-bands, which makes it particularly difficult to derive a high number of syllables for that segment. For ostensibly slow segments, there are sometimes high energy phonetic onsets that are strongly correlated across bands and form distinct peaks that are usually associated with syllable onsets; this effect tends to increase the measure past the transcribed rate.

Despite these limitations, *mrate* is clearly a better estimate of transcribed syllable rate than *enrate* was. As noted in Table 2, we have observed strong improvements in correlation, mean error, and error variance (standard deviation) with respect to the transcription-based standard. The sub-band component of *mrate* provided much of this improvement, as shown in the table in the sub-*mrate* row, but the complete measure is still significantly better.

Figure 3 shows an ROC curve for the detection of "fast" segments. For the purpose of this figure, we labeled as a false positive any segment that had been falsely classified as a "fast" (upper third in transcribed syllable rate, with the threshold determined from the training data). In this case as with the correlations, *mrate* is clearly much better than *enrate* for any choice of operating point.

Figure 4 displays the *mrate* ROC curve for the relaxed criterion, in which segments were only counted as false positives if they were strongly misclassified as fast (i.e., came from the bottom third of transcribed syllable rates).

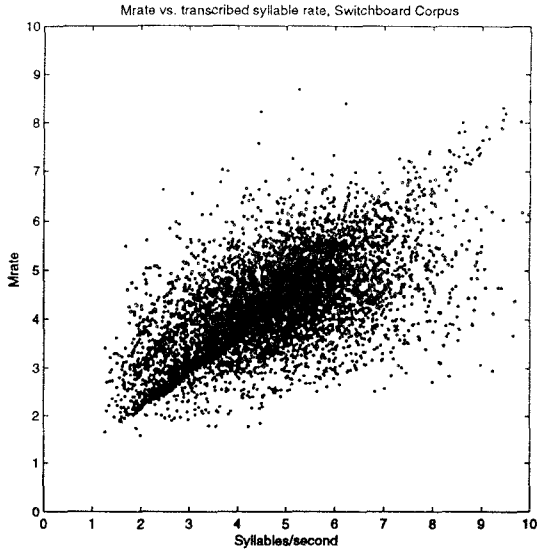

Figure 2: *mrate* versus transcribed syllable rate on roughly 3 hours of Switchboard between-pause segments taken from 4 hours of speech.

Similar curves have been observed for both criteria for the detection of unusually slow speech, but are omitted here for brevity.

Any particular choice of an *mrate* threshold corresponds to an operating point on the ROC curve. For *mrate* thresholds corresponding to the thirds of the data observed in the training set, classification results are shown in the confusion matrix of Table 3. Note that when 58% of the fast segments are detected, 13.1% of the slow segments are falsely detected as fast.

	mrate categories			
transcribed rate	slow	medium	fast	nsegs
slow	57.9%	29.0%	13.1%	214
medium	24.6%	43.8%	31.6%	272
fast	9.6%	32.4%	58.0%	386

Table 3: Confusion matrix between rate categories using transcribed rate versus *mrate*, with thresholds determined from 3 hours of data and results given for an independent 22 minute test set. Confusions are normalized for each of the transcribed categories so that percentages in each row sum to 100%. The matrix is augmented by a column giving the number of test set examples in each transcribed category. The uneven spread for categories chosen to split the training set into thirds shows that this particular test set has a higher proportion of "fast" segments than were found in the training set.

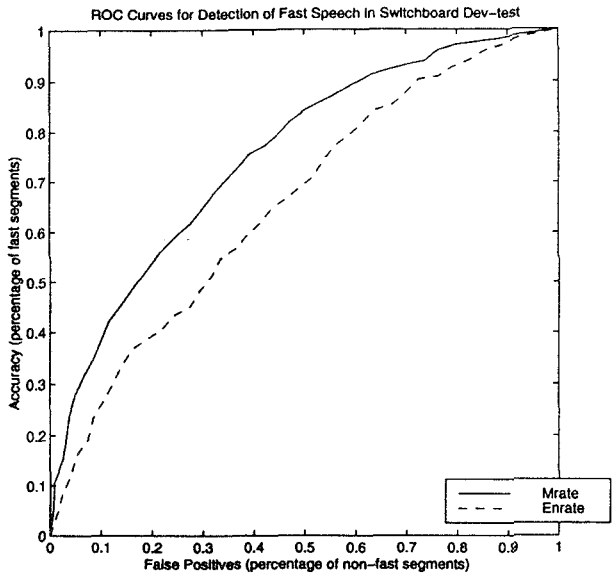

Figure 3: ROC curve for detection of fast between-pause segments in the test set, where "fast" is defined as the upper third of the training set distribution, according to manually transcribed syllable rate. The upper curve uses *mrate*, while the lower curve uses *enrate*.

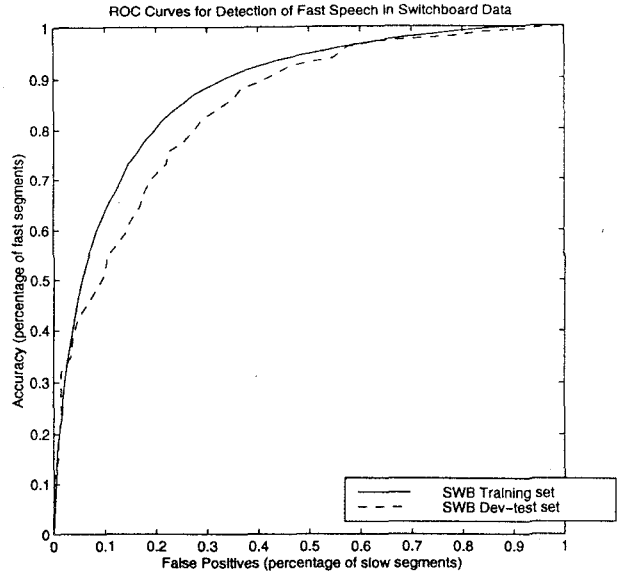

Figure 4: ROC curve for detection of fast between-pause segments in the test set. For this curve, segments with transcribed rates in the middle-third are not counted as either hits or false positives in the test. The figure shows the gap between training and test set curves.

5. CONCLUSIONS

We have developed *mrate*, a new method for estimating speaking rate from the acoustic signal. It can be used to detect unusually rapid or slow speech. For the conversational speech test set we evaluated, it assessed the transcribed syllable rate for a between-pause segment with an error whose standard deviation is 1.1. Once compensated for the difference in training set means, *mrate* was within 1 syllable per second of the transcribed rate 63% of the time, and within 2 syllables per second 88% of the time. Even with this spread, *mrate* may be useful for the detection of speech that is exceptionally rapid or slow. This has potential applications in the adaptation of duration models during speech recognition, as well as to dynamic pronunciation modeling.

The accuracy of the measure seemed to benefit greatly from the use of multiple estimators, some of which counted peaks and one of which incorporated a long term spectral estimate. The measure also seemed to benefit from the use of a pointwise cross-correlation between all pairs of spectral bands. This cross-sub-band measure was more correlated with syllable rate than any of the individual full-band measures we tried.

6. ACKNOWLEDGMENTS

We would especially like to thank Steve Greenberg for his suggestion to incorporate coherence between low frequency energy trajectories as part of a rate measure, and for the work by him and his team for generating the Switchboard syllabic/phonetic transcriptions. In addition, we continue to profit from discussions and suggestions from the entire Realization group at ICSI; in particular, this paper benefited greatly from the counsel of Nikki Mirghafori. This work was supported by a grant from the Center for Language and Speech Processing at The Johns Hopkins University, NSF SGER grant IRI-9713346, and NSF grant IRI-9712579.

7. REFERENCES

[1] Gold, B., and Rabiner, L., "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," *J. Acoust. Soc. Am.*, Vol. 46, No. 2, Pt. 2, pp. 442-448, Aug. 1969.

[2] Greenberg, S., "The Switchboard Transcription Project," in F. Jelinek, editor, *1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Reports*, chapter 6. Center for Language and Speech Processing, Johns Hopkins University, April 1997. Research Notes No. 24.

[3] Kitazawa, S., Ichikawa, H., Kobayashi, S., and Nishinuma, Y., "Extraction and Representation Rhythmic Components of Spontaneous Speech," *EUROSPEECH 1997*, pp. 641-644, Greece, 1997.

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

[5] Morgan, N., Fosler, E., and Mirghafori, N., "Speech Recognition using On-line Estimation of Speaking Rate", *EUROSPEECH 1997*, pp 2079-2082, Greece, 1997.

[6] 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.

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