FAST SPEAKERS IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION: ANALYSIS & ANTIDOTES

Nikki Mirghafori, Eric Fosler, and Nelson Morgan

International Computer Science Institute and
University of California, Berkeley
{nikki, fosler, morgan}@icsi.berkeley.edu

ABSTRACT

The performance of automatic speech recognizers (ASR) typically degrades for test speakers with "outlier" characteristics, for example, speakers with foreign accent and fast speaking rate. In this work, we concentrate on the latter. Consistent with other researchers, we have observed that for speakers with exceptionally high speaking rate, the word recognition error is significantly higher. We have investigated two possible causes for this effect. Inherent spectral differences may cause the extracted features for these outliers to be significantly different from that of normal speech. Also, due to phone omissions and duration reduction, the normal word-models may not be suitable for fast speech. Based on our exploratory experiments on TIMIT and WSJ corpora, we believe the spectral differences and duration reduction are both significant sources of the increased error. By adapting our MLP phonetic probability estimator to fast speech, and employing fast speaker word-models, we have been able to eliminate about 16% of the fast speaker word recognition errors.

1. INTRODUCTION

In a recent NIST WSJ evaluation (Nov 93) all participating systems had about 2-3 times higher word error rates on the two fastest speakers [4] (see Figure 1). In an earlier NIST Resource Management (RM) Sep 92 evaluation, this strong effect was also observed, as all participating systems had 2-4 times more error on the fastest (and one of the slowest[1]) speakers [5]. This observation naturally inspires the following question: "why do the ASR systems perform significantly worse on fast speakers?".

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 are inherently different from normal speech and these differences are 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 fast speakers often violate the phonemic durational constraints of the word-models (*durational errors*) or omit phones altogether (*deletion errors*). In the following sections, we describe our investigation of these two hypotheses using the TIMIT and WSJ corpora, and suggest corrective measures which give us about 16% relative improvement for fast speech.

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

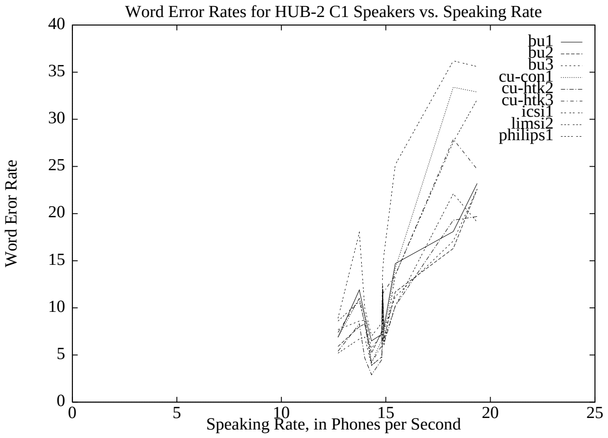

Figure 1: Rate of speech vs. word error rate for all participating cites in the WSJ0-93 Hub2 C1 evaluation. Each point represents one of the ten evaluation speakers. For WSJ0 training set $\mu_{ROS} = 13.24$ and $\sigma_{ROS} = 1.80$.

In our experiments, we use ICSI's hybrid HMM/MLP speech recognition system, which participated in the WSJ 93 (5K bigram task) and RM 92 NIST evaluations. As observed in Figure 1 and discussed in the work of Siegler [6], similar rate of speech (ROS) effects have been observed for mixture of Gaussian systems, so it is hoped that the conclusions of our work are useful in those systems as well.

2. ANALYSIS

In the following two sections, we discuss our investigation into the causes of higher error rate for fast speech.

2.1. SPECTRAL FEATURES

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 is useful for such multi-dimensional investigations.

In order to eliminate any word model effects (due to automatic labeling and alignment), we chose the hand-labeled TIMIT database and calculated the ROS for 4620 training sentences. The ROS for a particular sentence was calculated by dividing the number of non-silence transcribed phones by the non-silence duration of the sentence. For TIMIT training sentences, μ_{ROS} is 13.71 phones/second, and σ_{ROS} is 1.95 phones/second; the spread approximates a Gaussian distribution very well (Figure 2). For the female sentences $\mu_{ROS} = 13.43$ and $\sigma_{ROS} = 1.81$; For male sentences $\mu_{ROS} = 13.83$ and $\sigma_{ROS} = 1.99$. It is very interesting to note that the 3% relative difference in speaking rate between males and females is significant at $p < 0.001$ level! [2]

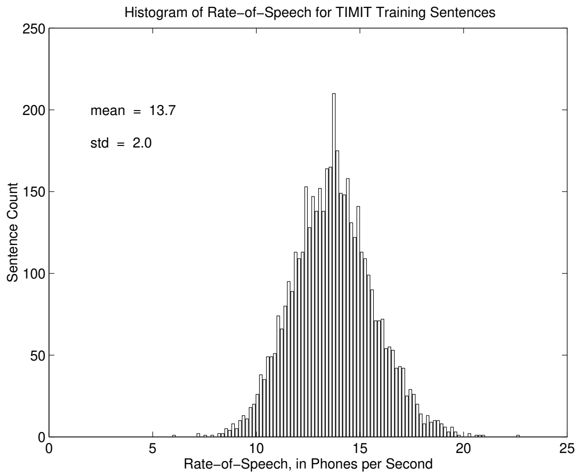

Figure 2: Histogram of rate of speech for training speakers of TIMIT.

We chose 400 sentences from the SX training set, 100 for each combination of $\{fastest, slowest\} * \{male, female\}$. Then we calculated the PLP12 & energy features and their deltas [1] (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) 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 extended our limited data by using a jack-knifing approach, by training on 90% of the data and testing on the remaining 10% for each of the ten possible such splits.

The mean classification accuracy for all phones on the tests was 73% (which is significantly higher than random) 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 makes particular sense in the light of psycho-acoustical studies that suggest diphthongs and glides are most affected by ROS variations [3]. The most difficult phones for speed discrimination were, unsurprisingly, the silence phones, closures, stops, and some fricatives.

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 frame error of the MLP phonetic probability estimator. In order to see whether there is any general correlation between ROS and the errors of the MLP, 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 (Figure 3). We see that for sentences which lie outside $\mu_{ROS} \pm \sigma_{ROS}$, the frame error is at least 2% higher. However, it is not clear how this frame error translates into word recognition error. We will attempt to answer this in section 3.1.

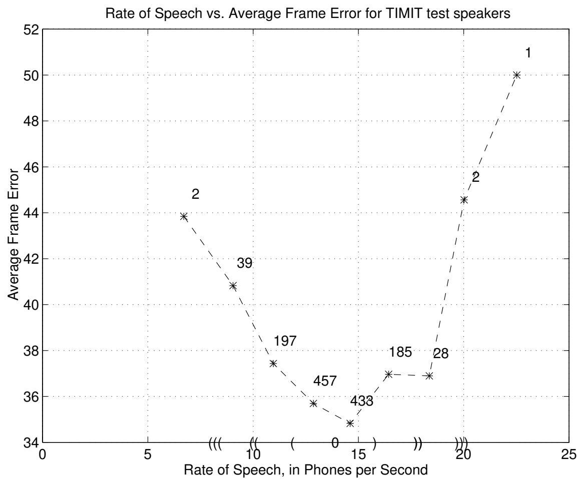

Figure 3: Rate of speech vs. MLP frame error for TIMIT test sentences. Each point represents the average error for a given ROS bin. The numbers on the graph denote the number of sentences in each bin.

2.2. WORD MODELS

The next question is whether the higher error rate is due to a mismatch with the word models. Our hypothesis is that the 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 n states.[3] 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 are measurable sources of error in our speech recognizer. However, since we had very limited

[2] Whether the information content per second is higher for male speakers is debatable, however.

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

hand transcribed data, we repeated this experiment on the TIMIT database. Similar to the analysis in 2.1, we divided the sentences into ROS bins, each $\frac{1}{2}\sigma_{ROS}$ wide. (Figure 4). There was almost no correlation between ROS and deletion errors alone (corr. coef. = -0.07)[4]. The correlation between ROS and durational errors was significantly higher at 0.84. Combining the deletion and duration errors, the correlation increases to 0.93 (Figure 4).

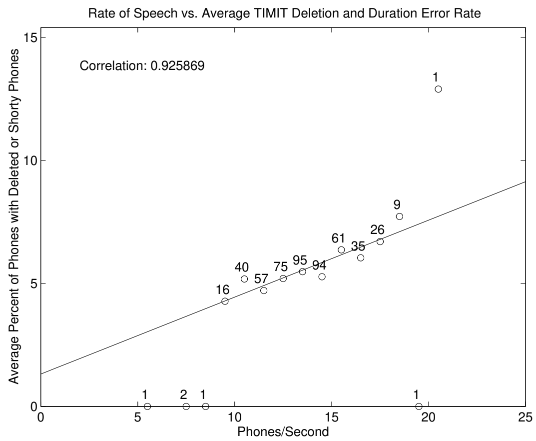

Figure 4: Rate of speech vs. average duration & deletion errors for TIMIT training sentences. The integers on the plot represent the number of sentences in each ROS bin.

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 3.2.

3. ANTIDOTES

In the following two sections, we discuss our experiments in trying to alleviate the higher error rates of fast speech.

3.1. ADAPTING THE MLP

Based on our observations in section 2.1, we decided to adapt our MLP phonetic estimator to fast speech. We chose the 5% fastest sentences (a total of 367) from the WSJ0 training corpus (C = ROS Cutoff = $\mu + 1.65\sigma = 16.17$). We adapted our 4000 hidden unit MLP, which was already trained on all of WSJ0, by retraining it on these fast sentences for three more epochs.

We tested this adapted net on the WSJ0-93 evaluation set. We looked at the word recognition error rate of sentences with $ROS > C$ (53 sentences) and $ROS < C$ (162 sentences). The "fast" sentences improved by 14%, while the "slow" sentences degraded by 10% relative to the baseline system.

3.2. CHANGING THE WORD MODELS

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 5.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. Our baseline WSJ0 recognizer[5] gives 16.1% word error for the WSJ0-93 evaluation set using these models.

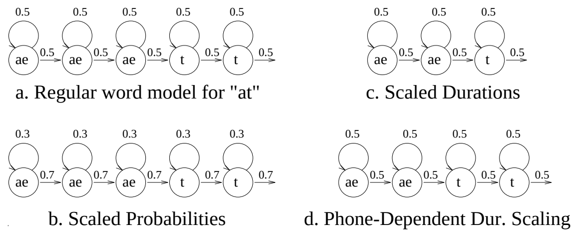

Figure 5: Examples of word models for "at"

In Figure 5.b, we show models where we scaled the probabilities of each HMM state to favor leaving rather than staying in the state. We found that for the sentences with $ROS > C$ of WSJ0-93 evaluation set, the exit probability x could be scaled as high as 0.9, with 15% relative improvement. The system overall performed best on *all* sentences at $x = 0.7$, giving 15.7% error, but assuming an ideal ROS detector the system would have improved to 15.0% error. Such a detector could be approximated by approaches discussed in [6], perhaps in combination with local detectors as described earlier in section 2.1.

An alternative would be to simply reduce the minimum phone durations. We tried this in both phone-independent (Figure 5.c) and phone-specific (Figure 5.d) duration scaling experiments. For the phone-independent models, experiments were conducted where 0.5 to 3 frames were subtracted from the average backoff trigram context-dependent duration of each phone. This resulted in an average of 0.25 to 1.5 state deletions in the minimum duration of phone models, causing a 5% relative decrease in error for fast sentences. When we reduced durations in a phone-specific manner by only reducing the average context-dependent durations of the vowels, the word error rate of fast sentences improved by 7% relative to the baseline system. In both cases, the overall recognition suffered slightly due to increase in slow speaker error.

Finally, we have introduced alternate pronunciations into our word models which represent the phone reduction and deletion effects often seen in fast speech [8, 9, 10, 2]. These pronunciations were generated by twenty surface-phonological rules 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. We feel that a phonological-rule based system for fast speech holds promise, and we plan to explore this avenue further in the future.

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

[5]Our baseline WSJ0 recognizer is a gender-independent system, with context-independent and one state per phone word models, and utilizes a 5K bigram grammar.

Word Error Rate for WSJ0-93 Evaluation Sentences			
Net	Lexicon	$ROS < C$	$ROS > C$
Base	Base	12.0	27.9
Adapt	Base	13.2	24.0 (14%)
Base	ScaledExitProb	13.6	23.7 (15%)
Base	ScaledDur	13.3	26.4 (5%)
Base	ScaledVowelDur	13.5	25.9 (7%)
Adapt	ScaledExitProb	14.8	23.3 (16%)

Table 1: The table shows all the results for the sentences with ROS above $C = \mu + 1.65\sigma = 16.17$ phones/sec and below C. The slow sentences' scores suffer from all the "ROS antidotes", while the fast sentences' error rates decrease. Percent improvement relative to the baseline system is shown in the parentheses. *Base Net* refers to our gender-independent, ROS independent MLP phonetic probability estimator, which is part of our WSJ baseline system. *Base Lexicon* is a context-independent with one state per phone lexicon of our WSJ baseline system. *ScaledExitProb* refers to word models with increased exit phone probability as discussed in section 3.2 and Figure 5.b. *ScaledDur* and *ScaledVowelDur* refer to reducing the minimum phone durations for the word models for all phones and for vowels only, respectively. These are also discussed in section 3.2 and shown in Figures 5.c and 5.d.

3.3. MERGING THE TWO SOLUTIONS

We combined the above approaches by using the phonetic probabilities from the adapted net and the ROS-tuned lexicon (Figure 5.b) for decoding (Table 1).

The merged system gives slight improvement (16% relative) over the system with scaled probabilities in the word models (15%) relative and the system with adapted MLP (14%). It is likely that to some extent each of the systems is compensating for similar variabilities of fast speech.

4. CONCLUSIONS

We have conducted a number of exploratory experiments to determine the likely sources of speech recognition errors due to unusual rates of speech (in particular, fast speakers). We believe 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. Another observable association has been between inappropriate word models for fast speech (due to exceptionally short phone duration or deletion) and recognition error rate.

We performed an adaptation of the MLP phonetic probability estimator for fast speech. This adapted net, reduced the error of the fast sentences by a relative 14%. Increasing the exit probability of the word models, which should alleviate *duration errors*, reduced the error on fastest sentences by a relative 15%. The merged system improved the word recognition error rate of fast speakers (i.e., speakers with $ROS > \mu + 1.65\sigma$) by 16% relative to the baseline system. In all cases, the error of the slower sentences was increased. Assuming an ideal ROS detector (an approximation of which is discussed in [6]), the overall error of our system on WSJ-93 evaluation set would be 14.9%, which is an improvement over 16.1% of our baseline system. More importantly, the ROS-tuned system is more robust 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 merged ROS system, however, reduces this error to 20%, effectively getting rid of 50% of the word errors. Now we face the challenge of implementing a reliable ROS detector and integrating it into our system.

Acknowledgments

Thanks to Hervé Bourlard, Steve Greenberg, Yochai Konig, Dan Jurafsky, and Gary Tajchman for their helpful comments and feedback, 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.

5. REFERENCES

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

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

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

[4] 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.

[5] 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.

[6] 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.

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

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

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

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