WHY IS ASR HARDER FOR FAST SPEECH AND WHAT CAN WE DO ABOUT IT?

Nikki Mirghafori, Eric Fosler, and Nelson Morgan

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

1. INTRODUCTION

It has been observed in various NIST evaluations (e.g. WSJ-Nov93 & RM-Sep92) that ASR systems typically have about 2-3 times higher word error rates on very fast speakers [2, 3]. This observation naturally inspires the following question: "why do ASR systems perform significantly worse on fast speech?"

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.

In our experiments, we use ICSI's hybrid HMM/MLP speech recognition system. Since similar rate of speech (ROS) effects have been observed for mixture of Gaussian systems [2, 3, 4], it is our hope that the conclusions of our work are useful in those systems as well.

2. ANALYSIS

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 hypothesize that we can train a classifier to distinguish between *fast* and *slow* phones based on the extracted features.

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/sec, and σ_{ROS} is 1.95 phones/sec; the spread approximates a Gaussian distribution very well. For the female sentences $\mu_{ROS} = 13.43$ and $\sigma_{ROS} = 1.81$; for male sentences $\mu_{ROS} = 13.83$ and $\sigma_{ROS} = 1.99$. We note that this 3% relative difference in speaking rate between males and females is significant at a $p < 0.001$ level![1]

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

We chose 400 sentences from the SX & SI training set, 100 for each combination of $\{fastest, slowest\} * \{male, female\}$. Then we calculated the PLP12 & energy features and their deltas (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.

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/, and /aw/ (mostly diphthongs and glides) the classification score was between 85-90%. This makes particular sense in the light of psychoacoustical studies that suggest diphthongs and glides are most affected by ROS variations [1]. 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 word recognition error rate for fast speakers. We tested this hypothesis by examining the frame error of the MLP phonetic probability estimator. We grouped the sentences in ROS bins each σ_{ROS} wide, and calculated the average frame error for each bin. We observed that the average frame error for sentences which lie outside $\mu_{ROS} \pm \sigma_{ROS}$ is at least 2% higher, and for sentences outside $\mu_{ROS} \pm 2\sigma_{ROS}$, the average error is at least 5% higher.

2.2. WORD MODELS

The next question is whether the higher error rate is also 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.

For the training sentences of TIMIT, we aligned each transcribed word with its corresponding word-model phonetic sequence, producing a *deletion error* score. Our word models (as with many other systems) have a minimum duration constraint, which require that each phone be repeated for n states. We calculated a *duration error* score which represents how often the transcribed phones were shorter than the minimum duration in the word model.

Similar to the analysis in 2.1, we divided the sentences into ROS bins, each $\frac{1}{2}\sigma_{ROS}$ wide and calculated average error for each bin. There was almost no correlation between ROS and deletion errors alone (ρ = -0.07). The correlation between ROS and durational errors was sig-

nificantly higher at 0.84. Combining the deletion and duration errors, the correlation increases to 0.93.

From these observations we conclude that the combination of unusually short sounds and deleted sounds are measurable sources of error in our speech recognizer.

3. ANTIDOTES

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 phones/sec). 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% (significant at a $p < 0.01$), 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 1.a, requires a minimal duration constraint. Our baseline WSJ0 recognizer[2] gives 16.1% word error for the WSJ0-93 evaluation set using these models.

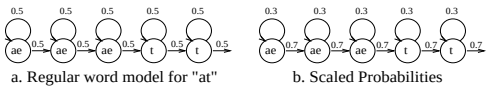

Figure 1: Examples of word models for "at"

In Figure 1.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 (signif. at a $p < 0.01$). The overall recognition suffered slightly due to increase in slow speaker error. Assuming an ideal ROS detector the overall system performance would improve to 15.0% error (signif. at a $p < 0.01$). Such a detector could be approximated by approaches discussed in [4], perhaps in combination with local detectors as described earlier in section 2.1.

We have also introduced alternate pronunciations into our word models which represent the phone reduction and deletion effects. The results of running with this lexicon and the adapted net were insignificantly worse than the base system. However, an error analysis of the results showed wide differences in error rate on a sentence-by-sentence basis between the two systems; the deletion lexicon removed up to 75% of the errors for some sentences, while for some others, it did worse. We feel that a phonological-rule based system for fast speech holds promise, and we plan to explore this avenue further in the future.

[2] 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.

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 1.b) for decoding. The merged system improves the error rate of the fast sentences ($C > ROS$) by 16% relative to the baseline system (significant at a $p < 0.01$).

4. CONCLUSIONS

We have conducted a number of exploratory experiments to determine the likely sources of speech recognition errors due to fast speech. 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.

A merged system that incorporates the adapted MLP and modified durational models 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. However, the error of the slower sentences was increased. Assuming an ideal ROS detector (an approximation of which is discussed in [4]), the overall error of our system on WSJ-93 evaluation set would be 14.9%, which is a significant improvement (at a $p < 0.01$) over the 16.1% of our baseline system. More importantly, the ROS-tuned system is potentially 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] Lindblom, B. Spectrographic Study of Vowel Reduction. *Journal of the Acoustical Society of America*, Vol 35, pp. 1773-1781, 1963.

[2] 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.

[3] 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.

[4] 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.