SPEECH DATA MODELING AT WS96 : THE QUESTIONABLE PARAMETER GROUP

Nelson Morgan

International Computer Science Institute/UC Berkeley

ABSTRACT

This is a summary of research conducted by the "Questionable Parameter Group" at the Johns Hopkins 1996 Summer Workshop. The focus in this group was on signal representation and acoustic modeling for conversational speech. In particular, long-time features, multi-stream analysis, and measures of speaking rate were studied. A four hour male-only subset of the Switchboard training materials was used to streamline development. A number of techniques led to improvements in word error rate over a baseline system using relatively "standard" features and single-stream probability estimation. Additionally, a preliminary signal processing measure was compared to phone and syllabic rates as indicated by manual transcription, and was found to be significantly correlated.

1. GROUP COMPOSITION AND GOALS

The work reported in this paper was the result of efforts by 8 individuals from 6 institutions; the author of this summary is responsible for any distortions of description, but the work itself was shared by the entire group. Members were: Hervé Bourlard from Faculté Polytechnique de Mons (FPMS) in Belgium; Jordan Cohen (group leader) from the Institute for Defense Analysis (IDA); Hynek Hermansky from the Oregon Graduate Institute (OGI); Nikki Mirghafori from the International Computer Science Institute (ICSI) and the University of California at Berkeley (UCB); Nelson Morgan from ICSI and UCB; Mark Ordowski from the Department of Defense; Christophe Ris from FPMS; and Sangita Tibrewala from OGI. All were interested in exploring two kinds of analyses for conversational speech recognition: the incorporation of long-term representations of speech in the estimation of acoustic probabilities, and the development of multi-stream approaches to this same estimation task. Additionally, we worked on estimating speaking rate from the acoustic signal.

2. EXPERIMENTS

To streamline experiments, we chose a limited subset of training and testing data, as well as simplified acoustic models. The training set for most experiments consisted of 4 hours of males from the workshop Switchboard training set, and the development test set was 240 utterances (also from males), or about 12 minutes of speech. The acoustic models were based on 56 monophones, each corresponding to a single density whose values were estimated with a multilayer perceptron (MLP) [1] that processed an input of some number of acoustic vectors (9 for the baseline system). Recognition was done using a lattice decoder developed at FPMS as part of the Speech Training and Recognition Unified Tool (STRUT) package. The decoder rescored lattices that were generated using HTK with a bigram grammar.

A baseline system incorporated these methods and an input to the MLP consisting of 9 frames of feature vectors, where each feature vector comprised 12th-order log RASTA PLP cepstral coefficients [2], delta cepstra for this same feature, delta-delta coefficients, and delta and delta-delta values for the zeroth cepstral coefficient This baseline system yielded a word error rate of 63.6% . By way of comparison, a full triphone-based HTK training with the same training, testing, and features yielded 60.0% error. The latter system differed in too many ways from the experimental approaches to be used as a comparative baseline, but the fact that the much simpler hybrid HMM/MLP system did not have too many more errors gave us confidence to use the hybrid as a baseline for these experiments.

2.1. Long-term processing

Conversational speech exhibits considerable temporal variability, so that short-term spectral analysis may not provide robust templates for statistical pattern recognition. Because of this, we explored features from much longer temporal regions than the usual 20-30 msec analysis window. While the MLP inherently permits the incorporation of larger time spans, explicit long-term features might provide additional robustness to short-term spectral variability.

There were two main experiments in this category:

1. "Chaf" features - A simple feature pair consisting of high frequency and low frequency delta log energy was

computed. A trajectory consisting of 25 of these pairs (computed over a 250 msec period centered at the "current" frame) was used as input to the MLP, along with 3 frames of mel spectra and their delta and delta-delta spectra. When probabilities from these features were used alone, they yielded poorer results than the standard features - we observed an error rate of 73.5%. However, when used in combination with the baseline system there was a small improvement in performance to 62.3%. Combination of the two probabilities will be discussed in a later section.

2. Low and high modulation spectra - Much as RASTA filters the trajectory of each log critical band energy (with a passband of roughly 1 to 12 Hz), we experimented with some more extreme filters. The lower one filtered the trajectory with a 2-6 Hz filter (roughly a syllable rate) and the higher one filtered the trajectory with an 8-14 Hz filter (roughly a phone rate). The resulting values were transformed into PLP cepstra as in the usual RASTA-PLP process. The lower band was used with a wide (21 frame) window at the input to the net, while the higher band was used with a narrow (5 frame) window. The final probabilities were combined with probabilities from the "standard" RASTA-PLP net. The new features were never better than simple RASTA, but combination always improved performance by about 1% to 62.4%.

2.2. Multiple stream processing

Another major theme in our work this summer was combining multiple streams of acoustic probabilities that were estimated from different observation streams. Streams were combined at the frame level and at the syllable level, although we only did some preliminary tests for the latter case. For recombination at the phone or syllable levels, we used a modified form of HMM decomposition [3]. We also experimented with the method of recombination - either summing log likelihoods or combining with another MLP. The former corresponds to an implicit assumption of stream independence, but the latter is more computationally demanding. When we were able to try the latter, performance was somewhat better.

Multiple stream processing was used for two different styles of experiment:

1. Long-term processing - for each of the experiments described in the previous section, we combined at the frame level by summing log likelihoods. Particularly for the "chaf" experiments, it is likely that an MLP recombination would be a better match to the feature choice, since the different features are highly correlated. For both "chaf" and high/low spectral modulation cases we are combining features with different temporal extent, so that a recombination at a longer-term level (such as the syllable) would probably have been better.

2. Multi-band processing - we also experimented extensively with combining subband likelihoods. Several experiments are described on our Web page, including work with 7 bands, but we describe the four band case here. The subbands were from 0-900, 800-1660, 1500-2550, and 2300-4000 Hz. Individual subband error rates were roughly 69%, while adding frame log likelihoods yielded an error rate of 61.4%. Using an MLP for recombination improved performance slightly (to 61.0%). Finally, recombining the multiband and full band likelihoods yielded a score of 59.4%, giving a raw error reduction of about 4%.

2.3. Speaking rate

In another set of experiments, we compared "enrate", a measure of the spectral moment for the low frequency energy envelope, to lexically-based measures of speaking rate. The latter were computed from manually-labeled phonetic transcriptions done at ICSI on 451 development test sentences. The correlation between the two measures was about .5. Handling pauses in a better way may be required to further improve the signal measure, since pause rate and articulatory rate are both included in the current measure.

3. SUMMARY AND CONCLUSIONS

During WS96, we began an ambitious set of experiments with long-term analysis features, probabilistic stream recombination, and signal measures of speaking rate. Preliminary results on a Switchboard development test subset were encouraging, with particularly strong results for subband recombination (see http://www.clsp.jhu.edu/ws96/ris/group.html for more details). Since the new features have different temporal properties than those used in the baseline, we expect stronger results once we realign the training set.

Finally, these experiments were made possible by the efforts of other contributors who provided extended computational tools in the form of SPERT hardware and software from ICSI and STRUT software from FPMS. We also thank our home institutions for supporting this effort.

4. REFERENCES

1. Morgan, N., and Bourlard, H., "Neural Networks for Statistical Recognition of Continuous Speech", *Proceedings of the IEEE*, pp. 742-770, May 1995.
2. Hermansky, H., and Morgan, N., "RASTA Processing of Speech", *IEEE Transactions on Speech and Audio Processing*, special issue on Robust Speech Recognition, vol.2, no.4, pp. 578-589, Oct. 1994.
3. Varga A.P. and Moore R.K., "Hidden Markov Model decomposition of speech and noise," *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 845-48, 1990.