AN INTEROPERABILITY STUDY OF SPEECH ENHANCEMENT AND SPEECH RECOGNITION SYSTEMS

Burhan Necioglu, Bryan George, George Shuttic

Signal Processing Center
The MITRE Corporation
McLean, VA 22102-3481 USA
email: {necioglu, bgeorge, gshuttic}@mitre.org

Ram Sundaram and Joe Picone

Institute for Signal and Information Processing
Mississippi State University
Mississippi State, MS 39762 USA
email: {sundaram, picone}@isip.mstate.edu

ABSTRACT

Speaker-independent automatic speech recognition (ASR) systems using Hidden Markov Modeling have evolved to the point where their performance and robustness are considered useful for military and industrial applications. At the same time, signal processing-based speech enhancement techniques have emerged that have clearly demonstrated their ability to deal with harsh background noise in narrowband communications environments. In remote information access applications, ASR systems will have to be interoperable with such speech enhancement techniques. It is thus critically important to study the effects of tandeming speech enhancement and ASR. This paper presents an initial study of these effects in the context of the recent DARPA SPeech In Noisy Environments evaluation, and suggests ways to improve the performance of integrated speech enhancement/ASR systems.

1. INTRODUCTION

Robust speech recognition is an attractive option to achieve automated, distributed access to information systems in military and industrial applications. However, medium to large-vocabulary continuous speech recognition is not currently practical for thin-client handheld platforms. “Server-based” speech recognition, whereby speech is transmitted via a wireless link to a central location, avoids the logistical issues associated with fielding equipment to support ASR, but introduces issues of bandwidth consumption and security.

Secure, narrowband voice communication is possible using speech coding techniques operating under 16 kbps. However, to varying degrees narrowband speech coders are subject to the assumption that source signals are speech from a single talker. As a result, speech coders often exhibit poor performance when presented with speech signals corrupted by noise.

Under DoD sponsorship, AT&T Research has designed a Harsh Environment Noise Pre-Processor (HENPP) algorithm [1], shown in Figure 1, to enhance speech in the presence of tactical ambient noise.1 The HENPP algorithm combines a short-term log spectral amplitude estimator with a soft-decision gain modification to eliminate musical noise effects, and a noise adaptation scheme capable of tracking non-stationary noise in the presence of speech. The HENPP algorithm has been integrated with the Federal Standard Mixed Excitation Linear Prediction (MELP) 2.4 kbps speech coder [2], and has been demonstrated to boost intelligibility of coded speech spoken in the presence of ambient noise [3].

Speech enhancement and speech coding algorithms are designed to optimize the goal of human-to-human communication, rather than human-machine communication. Before integrated human computer interface systems using speech enhancement and coding can be successfully deployed, it will be critically important to study and improve these components for interoperability with ASR systems. This paper presents the results of our participation in the recent DARPA SPeech In Noisy Environments

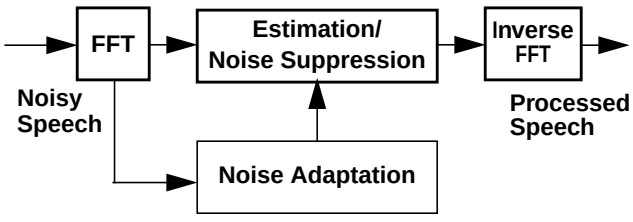

Figure 1: An overview of the noise preprocessor.

(SPINE) evaluation. In our system, the HENPP serves as a front-end signal processor. The ASR system is trained and tested using noisy speech as a baseline, then with speech processed with the HENPP. We analyze the resulting changes in recognition performance.

2. ASR SYSTEM OVERVIEW

The system used for the SPINE evaluations is a public domain cross-word context dependent HMM-based system. It consists of three primary components: the acoustic front-end, HMM parameter estimation module and a hierarchical single-pass Viterbi decoder.

The decoder [4] is based on a hierarchical implementation of the standard time-synchronous Viterbi search paradigm. The system uses a common front-end that transforms the input speech signal into mel-spaced cepstral coefficients appended with their first and second derivatives. The evaluation system used the front-end to generate 12 FFT-derived cepstral coefficients and log-energy. These features were computed using a 10 ms analysis frame and a 25 ms Hamming window. First and second derivative coefficients of the base features are appended to produce a thirty-nine dimensional feature vector. The features are made more robust to channel variations and noise by applying cepstral mean subtraction (CMS) using a mean computed over an entire conversation (side-based CMS).

Training for the SPINE evaluations was performed using an Expectation-Maximization based acoustic optimizer that used Baum-Welch algorithm for robust parameter estimation. The training algorithm supports continuous-density Gaussian mixture models with diagonal covariances. To overcome the problem of lack of training data for all the context-dependent models, the system uses maximum likelihood phonetic decision tree-based state-tying.

The evaluation system for SPINE, shown in Figure 2, was trained using conversation side-based CMS features for 10 hours of SPINE data. Initial training was performed with context-independent models that were iteratively trained from a single mixture to 32 mixtures. These models were then used to generate phone-level alignments. Context-dependent models were seeded from single mixture monophones and further training was done after state-tying. Mixture splitting was done using iterative splitting, terminating with 12-mixture word-internal models.

The SPINE lexicon and trigram language model (LM) were provided by Carnegie-Mellon University. A bigram LM was obtained by pruning all trigrams from the trigram LM. The lexicon used by the system had a vocabulary of 5226 words derived from the SPINE training data. The bigram LM had 5226 unigrams and 12511 bigrams. Recognition was performed in a single stage by doing a bigram decoding of the test data using word-internal models. All processing was performed at a real-time rate of 100x on a 600 MHz Pentium III processor.

The evaluation material consisted of a speech database previously collected [5] during the selection process for the 2400 bps Federal Standard vocoder.

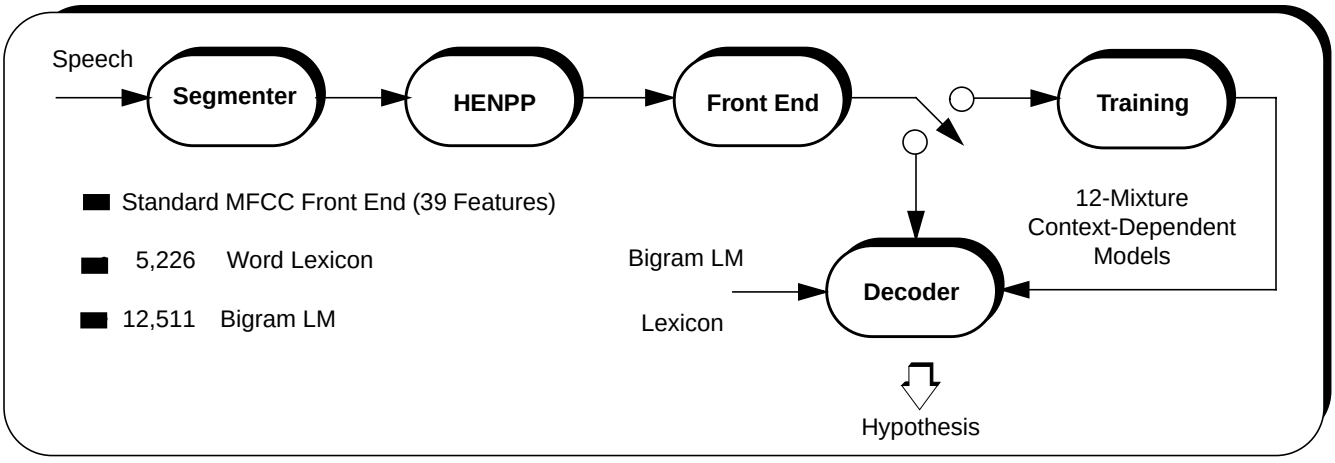

Figure 2: An overview of the integrated system used in the SPINE evaluation.

Conversations were recorded in sound booths between collaborating user pairs who were communicating through various channels and vocoders, and using different headsets while being subjected to pre-recorded noise types over loud-speakers. The training data consisted of approximately 7.5 hours of speech from 10 speaker pairs, with four different noise types: Aircraft carrier operations (AC), HMMWV, Office and Quiet. The evaluation data were approximately 10 hours long, with 20 speaker pairs, and two additional noise types: AWACS aircraft (E3A) and portable command and control shelter (MCE).

3. RESULTS AND ANALYSIS

Using the available training and evaluation data, two recognition experiments were performed: the baseline system with no noise cancellation, and the baseline system coupled with the HENPP front-end. In the latter experiment, both training and evaluation data were subjected to noise pre-processing. The recognition experiments did not utilize the noise type information for the conversation sides (an important detail). Conversation sides were pre-segmented using an energy based speech detector algorithm.

The error statistics from the evaluation, broken down according to the six provided noise types, are given in Table 1. Compared with the baseline system, the HENPP front-end increased the word error rate in almost all cases. Substitution errors were virtually the same except for MCE and Office, where the baseline system was better. Deletion errors were significantly better for the baseline system in all cases. In terms of insertion errors, the HENPP front-end either helped, or did not hurt performance, including the Quiet conversations. The HENPP front-end seemed to do better than the baseline for AC, performed virtually the same for E3A and HMMWV, and hurt performance for MCE, Office and Quiet conditions.

Since only the type of noise was given without any signal-to-noise ratio (SNR) information, it was not possible to directly gauge the effect of the HENPP front-end on recognition performance as a function of the noise level present in a speech segment. For this purpose, a blind SNR estimation algorithm was applied to the segmented conversation sides. Table 2 lists the statistics of the estimated SNR figures for all six noise types in the evaluation database.

Following the SNR estimation of the pre-segmented conversation sides in the evaluation data, a more detailed analysis was performed across the six noise types and four SNR ranges for each noise type. When considering the number of correctly recognized words, in 18 of the 24 cases (of noise type and SNR range), the baseline system performed significantly better, including some noisier cases such as AC. In the remaining 6, they were not significantly different. For substitution errors, the two systems were virtually identical in 23 of the 24 cases. The baseline was better in only one case: MCE. Considering

	AC	E3A	HMMWV	MCE	Office	Quiet	All Noise	All
Substitutions:								
Baseline	26.96	29.68	27.16	31.27	21.46	20.62	27.18	26.03
HENPP	27.61	30.14	25.99	33.57	23.07	21.96	28.20	27.10
Deletions:								
Baseline	27.48	27.18	15.58	20.85	17.90	17.57	21.90	21.13
HENPP	31.08	31.37	16.92	26.64	21.19	20.61	25.77	24.86
Insertions:								
Baseline	21.51	9.74	4.64	9.35	5.00	3.72	10.12	8.99
HENPP	14.82	6.24	5.21	6.16	4.06	3.03	7.23	6.49
Total Errors:								
Baseline	75.95	66.59	47.38	61.47	44.36	41.92	59.20	56.15
HENPP	73.52	67.75	48.12	66.37	48.32	45.60	61.20	58.45

Table 1: An analysis of performance (percent WER) by noise condition demonstrated that HENPP processing was effective for only one noise condition (the shaded cell), even though the system produced measurable improvements in SNR.

Condition	Avg	Min	Max
AC	25.4	12.6	35.2
E3A	23.5	12.5	34.6
HMMWV	24.6	11.2	36.0
MCE	27.8	17.9	36.2
Office	31.8	24.0	42.7
Quiet	32.6	24.2	37.0

Table 2: An analysis of the noise conditions by SNR (in dB).

deletion errors, the baseline performed better in 13 out of 24 cases, including some noisier cases (AC, E3A, MCE, and Office).

For the remaining 11 cases, the two systems performed almost identically. In terms of insertion errors, the HENPP front-end system performed better in 12 of the 24 cases (AC, E3A AWACS, and MCE ranges), and the two systems performed virtually the same for the remaining 12. Considering the total number of errors, or word accuracy, the baseline system performed better in 7 of the 24 cases including some noisier cases such as MCE. The two systems were virtually identical in 16 cases, including some noisier ones (AC, E3A and HMMWV). The HENPP front-end system performed better in one case only: AC. This was mostly due to the lower number of insertions. In summary, compared to the baseline, the HENPP front-end system only helped in reducing the insertion errors for half of the noise type/SNR range cases, and that by itself led to a better word accuracy only in one out of 24 cases (AC).

4. CONCLUSIONS

Our analysis of the recognition results presented provides two very interesting observations. The first is that there is generally a negative correlation between SNR and performance of the HENPP front-end system, i.e. the cleaner the speech, the more recognition performance is degraded by the HENPP algorithm. A comparison of clean speech with clean speech processed by the HENPP algorithm reveals that the HENPP algorithm by default is set to aggressively process noise, causing it to distort clean speech and degrading recognition performance. A second observation is that the HENPP algorithm induces a large number of deletion errors compared to the baseline, again mainly for clean speech. These errors are caused primarily by clipping initial and final consonants, which is an artifact of an aggressive approach to noise suppression.

This paper has presented an initial study about the implications of combining a speech enhancement preprocessor with a state-of-the-art recognition system in tactical noise environments. As expected, we note that speech enhancement, while effective at low SNRs, caused significant problems for ASR in less harsh environments. Our analysis indicates that recognition performance of an integrated speech enhancement/ASR system will improve considerably if speech enhancement is adjusted to reduce distortion of clean speech while maintaining noise suppression for low SNRs.

REFERENCES

[1] A. J. Accardi and R. V. Cox, "A Modular Approach to Speech Enhancement With An Application to Speech Coding", *Proceedings of the IEEE Int. Conf. on ASSP*, pp. 201–204, Phoenix, Arizona, USA, May 1999.

[2] A. McCree, *et al*, "A 2.4 kbit/s MELP coder candidate for the new U.S. Federal Standard", *Proceedings of the IEEE Int. Conf. on ASSP*, pp. 200–203, Atlanta, Georgia, USA, May 1996.

[3] J. S. Collura, *et al*, "The 1.2Kbps/2.4Kbps MELP speech coding suite with integrated noise preprocessing", *Proceedings of the IEEE Military Com. Conf.*, vol. 2, pp. 1449-1453, Atlantic City, NJ, USA, October 1999.

[4] N. Deshmukh, A. Ganapathiraju and J. Picone, "Hierarchical Search for Large Vocabulary Conversational Speech Recognition," *IEEE Signal Processing Magazine*, vol. 16, no. 5, pp. 84-107, September 1999.

[5] E.W. Kreamer and J.D. Tardelli, "Communicability Testing for Voice Coders," *Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1153-1156, Atlanta, Georgia, USA, May 1996.