SPEECH IN NOISY ENVIRONMENTS: ROBUST AUTOMATIC SEGMENTATION, FEATURE EXTRACTION, AND HYPOTHESIS COMBINATION

Rita Singh, Michael L. Seltzer, Bhiksha Raj[1] *and Richard M. Stern*
Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213 USA
1. Currently at Compaq Computer Corporation, Cambridge, MA 02142, USA

ABSTRACT

The first evaluation for Speech in Noisy Environments (SPINE1) was conducted by the Naval Research Labs (NRL) in August, 2000. The purpose of the evaluation was to test existing core speech recognition technologies for speech in the presence of varying types and levels of noise. In this case the noises were taken from military settings. Among the strategies used by Carnegie Mellon University's successful systems designed for this task were session-adaptive segmentation, robust mel-scale filtering for the computation of cepstra, the use of parallel front-end features and noise-compensation algorithms, and parallel hypotheses combination through word-graphs. This paper describes the motivations behind the design decisions taken for these components, supported by observations and experiments.

1. INTRODUCTION

The first "Speech in Noisy Environments" (SPINE1) evaluation was conducted by the Naval Research Laboratories (NRL) in August, 2000. The purpose of the evaluation was to provide impetus to the design of algorithms and strategies which improve the performance of speech recognition systems in the presence of varied noises. The task consisted of recognizing approximately nine hours of speech from battleship games with realistic military noises playing in the background. Approximately eight hours of speech recorded under similar conditions were provided for training the systems. The training data had fewer speakers and noise types than the evaluation data. The signal-to-noise ratio (SNR) for both the training and test data varied from 5 dB to 20 dB.

The main feature of the evaluation was the variety of the background noises occurring in the background, and the mode of recording. The noises included aircraft and aircraft carrier sounds, with additional tones, pings, and background speech. Continuous recordings were made of signals from a communication line. The line was activated at approximately the time at which an utterance began, and was deactivated at the end of the utterance. The recorded signal consisted of a continuous background signal of noise produced by the recording equipment, with intermittent recordings of the speech and noise communicated through the channel. As a result, there were two distinct levels and types of noise in each recording, the first consisting of the recording-equipment noise, and the second consisting of the external military noises being used to corrupt the signal. Additionally, some segments of the recording exhibited *dropouts*, where short segments of speech within an utterance are deleted. Figure 1 shows a segment of the SPINE1 data to illustrate some of these features of the recording.

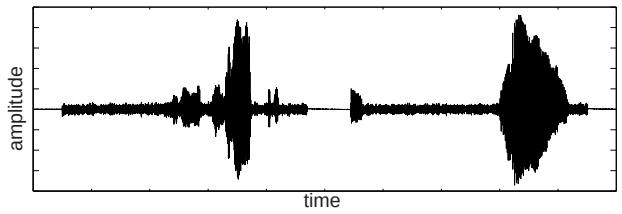

Fig. 1 A typical segment of speech from the SPINE1 data. Note the abrupt onset and termination of the signal, leading to two distinct levels of background noise at the beginning and the end of the utterance. The low-energy region in the middle of the utterance is a dropout occurring in the middle of a word. Such recordings are likely when the speakers are using a push-button recording setup.

Carnegie Mellon University (CMU) based its approach to this task on the hypothesis that for robust speech recognition it is preferable to extract multiple cues directly from the speech signal, rather than learn about the corrupting noises and compensate for them. This paper describes the key elements of two systems that were designed specifically to test this hypothesis.

In the CMU primary system, multiple feature representations were extracted from the uncompensated noisy speech and used to generate parallel recognition outputs. These outputs were then combined to generate the final recognition hypothesis. In the CMU secondary system, multiple compensation algorithms were applied to a single feature stream to generate multiple recognition outputs, which were then combined to obtain the final hypothesis. The principle of focusing on speech, rather than on the noise, was also central to the design of a session-adaptive segmentation strategy.

In the following section we describe the segmentation strategy. In Section 3 we describe the features and compensation algorithms used in the two systems, including specifically the use of wide-bandwidth mel filters in the computation of MFCCs for noise robustness. In Section 4 we describe the algorithm used to combine parallel hypotheses. In Section 5 we present recognition results obtained with the two systems. Finally, in Section 6, we present our conclusions.

2. SEGMENTATION

As described in the previous section, multiple noise levels were present in most SPINE1 recordings. In addition, the type and level of these noises varied from recording to recording, and even within a single recording. In this scenario, energy-based segmentation may fail to separate noise events from speech events, since a switch from one noise type to another is indistinguishable from a switch from noise to speech, especially if the dynamic ranges of noise and speech are variable. Under the circumstances, it may be necessary to utilize the characteristics of speech, rather than those of the noise, to segment the signal. We therefore used a two-class classifier- based segmenter, where the two classes were *speech* and *not speech* respectively. While the speech class referred to all sections of the signal which corresponded to actual speech, the not-speech class referred to all signal segments that

corresponded to all events other than speech. The latter class did not distinguish between different kinds of noises, or even between noise and silence. To train the speech class, speech segments were chosen after Viterbi alignment of the data to obtain word boundaries. All other data were used to train the non-speech class.

Although the class distributions are fixed for such a classifier once it is trained, it is important to note that the optimal *a priori* probability associated with each of the two classes, and therefore the optimal decision boundary that minimizes the classification error, varies with the utterance being segmented. It is dependent on the amount of data present from each class and its closeness to each class (which in turn is dependent on the type and level of noise present in the data). For each utterance being segmented, therefore, it is necessary to automatically determine the optimal decision boundary for the segmenter.

For this two class case, a Bayesian classifier can be represented as a discriminant function:

$$D(X) = \log\left(\frac{P(X|not\ speech)}{P(X|speech)}\right) + T = L(X) + T \qquad (1)$$

where X represents feature vectors from the segment of speech being classified, $L(X)$ is the difference in log likelihoods of not-speech and speech distributions, and T is a threshold, given by

$$T = \log\left(\frac{P(not\ speech)}{P(speech)}\right) \qquad (2)$$

Whenever $D(X)$ is positive, the segment X is identified as non-speech and when $D(X)$ is negative it is identified as a speech segment. T is a controllable threshold which can be used to optimize classification on the given utterance to be segmented.

The distribution of $L(X)$ usually exhibits two distinct modes, one representing the values of $L(X)$ when X belongs to speech, and the other mode representing the values of $L(X)$ when X belongs to not-speech. The precise positions of these modes vary from utterance to utterance, due to the fact that the distributions in the classifier may not be truly representative of the utterance being segmented. The bimodal characteristic of the distribution of $L(X)$, however, is ubiquitous.

Figure 2 shows the histograms of $L(X)$ for two typical recordings in the SPINE1 data, where $L(X)$ was computed on sets of feature vectors obtained from 0.5-second windows of speech. In each case the distribution is bimodal where the flatter mode comes from speech regions, and the sharper mode comes from non-speech regions. The point of inflection between the two modes identifies the point where X changes from being predominantly

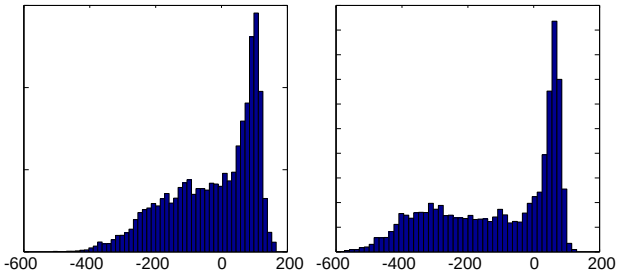

Fig. 2 Histogram of the difference in log-likelihoods of *not speech* and *speech* for two typical recordings in the SPINE1 data. In each case the distribution is bimodal where the flatter mode represents speech regions, and the sharper mode represents non-speech regions.

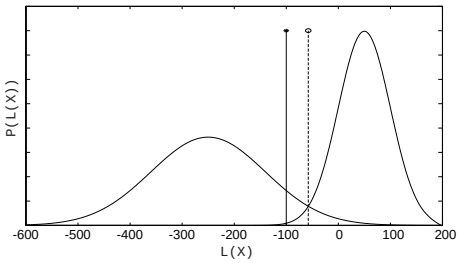

Fig. 3 Schematic representation of the bimodal distribution of $L(X)$. The optimal threshold, represented by the dotted line, is shifted toward the left to give a more conservative estimate of the threshold T.

speech to predominantly noise (left to right). This point is therefore the optimal value of T in the absence of other considerations.

In the SPINE1 data, however, the cost of identifying a noise segment as speech and hypothesizing words in it was observed to be typically higher than the cost of identifying parts of speech segments as noise and clipping them. This was because the number of insertions in the former case was greater than the deletions enforced in the latter case. In such situations the optimal value of T must be adjusted to increase the probability of identifying segments of the signal as noise, at the cost of erroneously clipping speech. Figure 3 shows a schematic diagram of the bimodal distributions observed in the histograms shown in Figure 2. The inflection point is the point at which the two modes cross over. For conservative segmentation of SPINE1 data, the optimal threshold T was shifted from this point to lie at the point identified by the reflection of the rightmost point of the noise mode across the position of the noise peak.

In the SPINE1 evaluation the CMU systems were observed to have the lowest number of "gap" insertions among all systems. Gap insertions are words that are hypothesized in non-speech segments. The gap insertions in the CMU systems were also observed to be relatively less sensitive to the noise type.

3. SIGNAL PROCESSING

3.1. CMU primary system

The CMU primary system used three parallel feature representations:

1. Wide-filter mel frequency cepstral coefficients (WMFCC)
2. Perceptual Linear Prediction cepstra (PLP) [1]
3. WMFCCs of speech that is low-pass filtered to 5 kHz.

Wide-filter mel frequency cepstral coefficients are a variant of conventional mel frequency cepstral coefficients (MFCC), and were designed to improve the noise robustness of the feature. In the computation of standard mel-frequency cepstra for speech, the speech signal is filtered using a fixed number of triangular filters. The filter index of each of the filters is related to its peak frequency by the mel curve. The filters are designed to have a 50% overlap, such that the peak of any filter coincides with the trailing edge of the earlier filter and the leading edge of the subsequent filters. Figure 4a shows a conventional band of mel filters with the 50% overlap. In the case of wide-filter mel frequency cepstra (WMFCC) the filter overlap is increased to 75%, instead of the conventional 50%. The peak frequencies of the filters do not change. Figure 4b shows a band of wide mel filters.

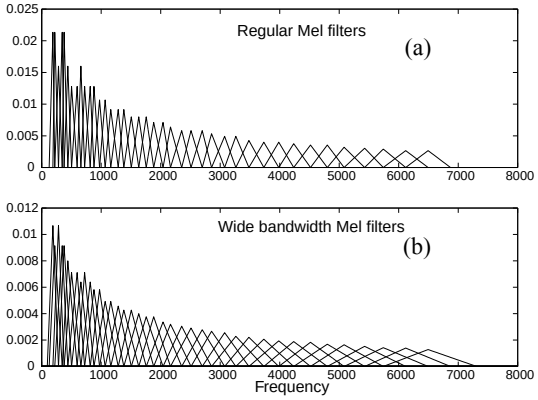

Fig. 4 a) A bank of 40 regular mel filters from with peak frequencies ranging from 130 Hz to 6800 Hz. Adjacent filters overlap by 50%. b) A bank of 40 wide mel filters with identical peak frequencies. Adjacent filters overlap by 75%.

Figure 5 shows the average distortion introduced by noise in the filter outputs for standard and wide mel filters. The distortion at the output of any filter is the energy in the error between the output of the filter when the input is clean speech, and the output when the input is noisy speech. We see from this figure that at any noise level (SNR) the average distortion is lower for the wide mel filters than for standard mel filters. As a result, WMFCCs show less variation with increasing noise than regular MFCCs. WMFCCs were also observed to improve WERs by a relative 5-7% compared to MFCCs for pilot experiments using SPINE1 training data.

3.2. CMU secondary system

In the CMU secondary system only a single feature representation (WMFCC) was used. The noisy signal was processed with four compensation algorithms: CDCN [2], VTS [3], KLT-based noise compensation [4], and SVD based noise compensation [5]. Of these, CDCN was performed on the noisy WMFCCs, VTS was performed on the (wide filter) log mel spectra of the noisy signal, and KLT and SVD were performed directly on the signal.

4. COMBINATION OF PARALLEL HYPOTHESES

The word hypotheses obtained from parallel systems were combined into a word graph. Initially each word in each of the hypotheses was represented by a node in the graph. The acoustic score of that word (from that hypothesis), was associated with the

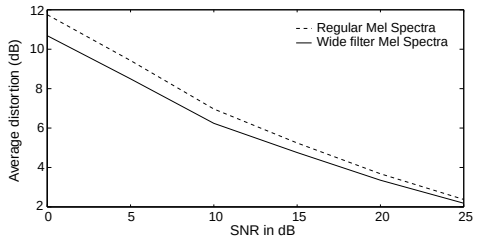

Fig. 5 Average distortion in mel filter output as a function of SNR for regular and wide bandwidth mel filters. The average distortion is seen to be lower for the wide mel filters.

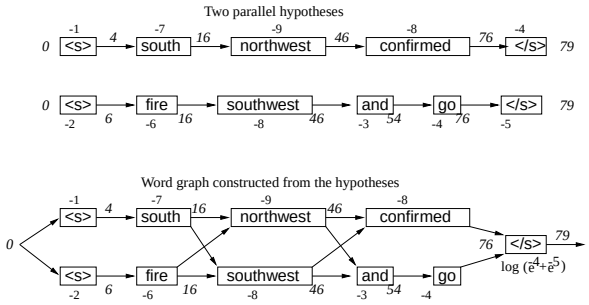

Fig. 6 Example of word graph construction from two parallel hypotheses. The upper portion of the figure shows two parallel hypotheses. The acoustic likelihoods of words are shown above and below the words. The italicized numbers identify the transition frames. Additional transitions are permitted across the hypotheses when both hypotheses have word transitions at the same instant. Since the final words in both hypotheses are identical and hypothesized between identical time instants, they are merged into a single node whose acoustic likelihood is a combination of the likelihoods of the original words.

node. In the next step all nodes representing identical words hypothesized between the same time instants were collapsed into a single node. Finally, links were formed between all node pairs where the word-end time of one node and the word-begin time of the next node were within 30 ms of each other. Figure 6 illustrates the formation of a word graph.

After the word graph was constructed in this manner, a standard language model was used to score the paths through the graph and the best path was obtained as the final hypothesis. Note that combining lattices generated during recognition in this manner would have provided much better resulting paths. However, this was not implemented for the current evaluation.

5. RECOGNITON RESULTS

The CMU Sphinx-3 speech recognition system was used in both systems. All acoustic models were continuous density 3-state context-dependent triphone HMMs with 2600 tied states, each modeled by a mixture of 8 Gaussians. In the primary system hypotheses generation was done in a series of steps. Recognition was performed with each of the three features and their hypotheses were combined. The combined hypothesis was used to perform two passes of MLLR adaptation of the acoustic models for each of the three parallel features. The recognition hypotheses of the adapted systems were combined and used to retrain the three parallel feature systems in an unsupervised manner. The retraining was aimed at shifting the systems out of any local optimum that the MLLR adaptation might have induced. The combined output of the retrained system was then used to perform a final pass of MLLR adaptation of the acoustic models, and their hypotheses were combined to give the final hypothesis for the primary system. Hypothesis combination at any stage was cumulative over all existing hypotheses till that stage. Table 1 shows the word error rates obtained with each of the parallel features at various stages of the hypothesis generation, as well as the combined hypothesis at each stage.

In the CMU secondary system the test data were compensated using each of four compensation schemes (CDCN [2], VTS [3], KLT[4], and SVD[5]), and WMFCCs were derived from each of them. The four hypotheses obtained from these four sets of

	First Pass	MLLR Adapt1	MLLR Adapt2	Retrain	MLLR
WMFC	35.1	33.3	32.9	30.6	29.9
PLP	38.0	34.8	34.7	31.6	32.1
LMFC	47.4	40.1	38.7	35.4	35.0
Comb.	32.8	30.2	28.4	27.3	26.5

Table 1: WERs at various stages of hypothesis generation in the CMU primary system. LMFC refers to low-pass filtered WMFCCs, and Comb. refers to the combined hypothesis.

WMFCCs were combined to obtain the final hypothesis. Table 2 shows the word error rates obtained with each of the compensation systems, as well as the final combined hypothesis. Additional passes of MLLR were not observed to result in any improvement of WER in the secondary system in pilot experiments performed on the training data. They were therefore not performed in this system.

CDCN	VTS	KLT	SVD	Comb
31.1	32.2	33.8	33.4	29.3

Table 2: WERs from the various compensation methods used in the CMU secondary system, and from the final combined hypothesis.

6. DISCUSSION AND CONCLUSIONS

When noise types and conditions vary significantly within the environment in which recognition is to be performed, and the environment is open to new and unknown noise conditions, no single noise compensation algorithm can be expected to always perform effectively. This is especially true if the compensation algorithm has been designed for a specific set of noise conditions or with a specific underlying noise model assumption such as linearity of the corrupting channel and/or stationarity of additive noise.

It is known that as the environment becomes more and more noisy, humans rely on more and more cues from the speech signal [6], sometimes even relying on other cues like lip movement and facial expression. Similarly, to perform speech recognition in unknown environmental noise conditions, we believe it is better to analyze the speech signal from many different perspectives in order to extract a greater number of informative cues about the speech itself. This differs from conventional noise compensation techniques which concentrate on extracting information about the noise in the environment. The value of utilizing parallel representations of information at various levels of a classification task is well known. Parallel representations of information have been successfully used in various fields, including machine learning [7] and speech recognition (*e.g.* [8,9,10]).

We designed two different speech recognition systems to verify this approach to noise robustness and compare it with the conventional approach. The CMU primary system was designed to extract more cues directly from the noisy speech signal, attempting to get information from multiple perspectives. The "perspectives" were equated to different parametrizations of the speech signal. In contrast, the CMU secondary system was designed to improve robustness through many parallel noise compensation algorithms. In this system, several noise compensation algorithms were applied in parallel to a single feature set or "perspective" of the speech signal. A comparison of the final WERs of the two systems showed that the parallel feature combination strategy was more robust to noise than the parallel compensation strategy.

ACKNOWLEDGEMENTS

The authors thank Pierre Ponce for his help in implementing the PLP features. This research was sponsored by the Department of the Navy, Naval Research Laboratory under Grant No. N00014-93-1-2005. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES

1. Hermansky, H., "Perceptual Linear Predictive (PLP) analysis of Speech", *Journal of the Acoustic Society of America*, vol. 87, no. 4, pp 1738-1752, 1990

2. Acero, A., *Acoustic and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, 1993

3. Moreno, P. J., *Speech Recognition in Noisy Environments*, Ph.D. Dissertation, Carnegie Mellon University, May 1996

4. Therrien, C. W., *Discrete Random Signals and Statistical Signal Processing*, Prentice Hall Inc., New Jersey, 1992

5. Hermus, K.,Wambacq, P., Compernolle, D. K., "Fully Adaptive SVD-Based Noise Removal for Robust Speech Recognition," *Proceedings of Robust Methods for Speech Recognition in Adverse Conditions*, pp. 223-226, Tampere, Finland, May 1999

6. Denes, P.D., Pinson, E. N., *The Speech Chain: The physics and biology of spoken language*, W. H. Freeman and Company, New York, 1993

7. Jordan, M. I., Jacobs, R. A., "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 6, pp. 181-214, 1994

8. Dupont, S., Bourlard, H., Ris, C., "Robust speech recognition based on multi-stream features", *ESCA-NATO workshop on robust speech recognition for unknown communication channels*, Pont-a-Mousson, France, 1997

9. Halberstadt, A. K., Glass, J. R., "Heterogeneous measurements and multiple classifiers for speech recognition", *Proc. ICSLP98*, Sydney, Australia, 1998

10. Janin, A., Ellis, D., Morgan, N., "Multi-stream speech recognition: Ready for prime time?," *Proc. Eurospeech99*, Budapest, 1999