SPECTRAL ESTIMATION FOR NOISE ROBUST SPEECH RECOGNITION

Adoram Erell and Mitch Weintraub

SRI International

ABSTRACT

We present results on the recognition accuracy of a continuous speech, speaker independent HMM recognition system that incorporates a novel noise reduction algorithm. The algorithm is a minimum mean square error estimation tailored for a filter-bank front-end. It introduces a significant improvement over similar published algorithms by incorporating a better statistical model for the filter-bank log-energies, and by attempting to jointly estimate the log-energies vector rather than individual components. The algorithm was tested with SRI's recognizer trained on the official speaker-independent "Resource management task" clean speech database. When tested with additive white gaussian noise, the noise reduction achieved by the algorithm is equivalent to a 13 dB SNR improvement. When tested with desktop microphone recordings, the error rate at 13 dB SNR is only 40% higher than that with close-talking microphone at 31 dB SNR.

I. INTRODUCTION

Speech recognition systems are very sensitive to differences between the testing and training conditions. In particular, systems that are trained on high quality speech degrade drastically in noisy environments. One of several commonly used methods for handling this problem is to supplement the acoustic front-end of the recognizer with a statistical estimator. This paper introduces a novel estimation algorithm for a filter-bank based front-end, and describes recognition experiments with noisy speech.

Estimation algorithms that have been used in filter-bank based systems can be classified by their estimation method — minimum mean square error (MMSE) versus subtraction — and by the features they estimate — DFT coefficients versus the filter-bank output energies. Boll [1] and Macaulay and Malpass [2] used spectral subtraction; Ephraim and Malah [3],[4] and Porter and Boll [5] used MMSE estimation of various functionals of the DFT coefficients; Van Compernolle used filter energy subtraction [6] and MMSE estimation of filter-bank log energies [7]. The latter MMSE algorithm lacks, however, the degree of precision that has been incorporated by Porter and Boll in their statistical modeling and was not compared to their algorithm.

A common deficiency of all of the above algorithms is that they estimate different frequency channels (whether DFT coefficients or filter output energies) independently. However, for a speech recognition system which is based on a distance metric, whether for template matching or vector quantization, the estimation should strive to minimize the average distortion as measured by the distance metric. For euclidean distance this criterion yields the following estimator

$$\hat{\vec{S}} = \int \vec{S}\, P(\vec{S} \mid \vec{S'})\, d\vec{S} \qquad (1)$$

where $\vec{S}$ is the clean speech feature vector and $P(\vec{S} \mid \vec{S'})$ is the aposteriori probability of the clean speech given the noisy observation. This estimator is considerably more complex than the independent MMSE of individual components (denoted by S_k),

$$\hat{S}_k = \int S_k\, P(S_k \mid S'_k)\, dS_k \qquad (2)$$

since Eq.(1) involves the estimation of multidimensional probability distributions and multidimensional integrations, whereas Eq.(2) is a relatively simple one dimensional integral. Both Eqs.(1) and (2) can in principle be computed using Bayes's rule, which than requires the conditioned probability $P(\vec{S'}|\vec{S})$ of the noisy observation $\vec{S'}$ given that the clean speech was $\vec{S}$, and the clean speech probability distribution $P(\vec{S})$.

Most HMM systems use probability densities or distance metrics in a transformed domain. For example, our recognizer uses a weighted euclidean distance on the cepstral transform of the filter-bank energies. Since the optimal estimation criterion cannot be easily satisfied for such a metric, the practical question is which features and what computationally feasible estimation scheme will best approximate the optimal estimator. We argue that the filter-bank log energies are more attractive to estimate relative to either the DFT or the cepstral coefficients. They are more attractive than DFT coefficients since (a) the euclidean distance between filter-bank energies vectors is a better approximation to the cepstral distance than a euclidean distance between any functional of the DFT coefficients, and (b), the estimation of typically 20-25 filter-bank energies is computationally easier than the estimation of typically 200 DFT coefficients. They are more attractive than the cepstral coefficients since the conditioned probability $P(\vec{S'}|\vec{S})$ can be modeled accurately for gaussian type noise in the frequency domain but not in the cepstral domain.

In the present work we achieved three objectives. First, we derived an MMSE estimator for filter-bank log-energies based on a more accurate statistical model than the one derived by Van Compernolle [7], and compared its performance to that achieved with the DFT estimator of Porter and Boll [5]. Second, we improved over the independent MMSE estimation of individual filter energies by computing an approximation to the minimum-distortion estimator, Eq.(1). Third, the estimation algorithm was evaluated with SRI's DECIPHER continuous speech recognition system [8] on the NBS "Resource management task" speech database [9] with both additive white gaussian noise, and with desktop microphone recordings.

II. ESTIMATION OF FILTER LOG-ENERGIES

A. MMSE OF INDIVIDUAL FILTER LOG-ENERGIES

The MMSE estimator given by Eq.(2) can be computed using Bayes's rule as follows:

$$\hat{S}_k = \frac{\int S_k \, P(S'_k \mid S_k) \, P(S_k) \, dS_k}{\int P(S'_k \mid S_k) \, P(S_k) \, dS_k} \tag{3}$$

where S_k is the clean speech filter log-energy and S'_k is the observed noisy value. The clean speech probability distribution $P(S_k)$ was estimated in our implementation from speech data. The conditioned probability $P(S'_k|S_k)$ was modeled as follows.

The filter output energy (E_k) is computed by a weighted sum of squared DFT coefficients. For additive, gaussian noise, the DFT coefficients of the noise are approximately independent, gaussian random variables. Approximating the weighted sum by a non weighted sum of M coefficients, and assuming that the noise spectral power is uniform over the frequency range spanned by any single filter, the noisy filter energy E'_k is given by

$$E'_k = \sum_i | DFT_s(i) + DFT_n(i) |^2 \tag{4}$$

where the subscripts s and n correspond to speech and noise. Since the noise spectral power is assumed to be uniform within the range of summation, both $Re\{DFT_n\}$ and $Im\{DFT_n\}$ are gaussian random variables with sigma given by

$$\sigma^2 = \frac{N_k}{2M} \tag{5}$$

where Nk is the expected value of the noise filter energy. Under these conditions the random variable E'_k/σ^2 will obey a probability distribution of non central chi square with 2M degrees of freedom and non central parameter λ [10], so that

$$P\left(\frac{E'_k}{N_k} \mid \frac{E_k}{N_k}\right) = 2M \cdot \chi^2_{N.C.}\left(\frac{2ME'_k}{N_k}, 2M, \lambda\right) \tag{6}$$

$$\lambda = \sum_i \frac{|DFT(i)|^2}{\sigma^2} = \frac{2ME_k}{N_k} \tag{7}$$

and the probability of the log-energy can then be easily computed.

To account for correlations between DFT coefficients (introduced for noise by the hamming window), we relaxed the parameter M to fit the above model to simulated distributions with white gaussian noise. The modeled conditioned probability $P(S'_k|S_k)$ and the estimated clean speech probability distribution $P(S_k)$ were then used to compute the MMSE estimator of individual filter log energies.

B. APPROXIMATE MINIMUM-DISTORTION JOINT ESTIMATION OF FILTER LOG ENERGIES

To improve over the individual components MMSE estimator we approximated the joint estimator, Eq.(1), by the following method: Eq.(1) can be computed with Bayes' rule, similarly to Eq.(3), with the vector S replacing the components S_k. The conditioned probability P(S'|S) can then be modeled simply as the product of the marginal probabilities:

$$P(\vec{S}' \mid \vec{S}) = \prod_k P(S'_k \mid S_k) \tag{8}$$

This approximation is fairly good for nonoverlapping filters, since gaussian noise is uncorrelated in the frequency domain and the value of a given noisy filter energy S'_k is indeed dependent only on the clean energy S_k and on the noise level in that frequency band. For overlapping filters, such as in our system, the approximation in Eq.(8) is not as good as for nonoverlapping ones, but is still quite reasonable. The clean speech probability distribution P(S), on the other hand, cannot be represented in the frequency domain as a product of the marginal probabilities. In fact, if it could have been represented as such, the joint estimate would have been reduced to MMSE of individual components. However, one can improve over the single product model of $P(\vec{S})$ by a sum of such products:

$$P(\vec{S}) = \sum_n C_n \prod_k P_n(S_k) \tag{9}$$

The acoustic space can be partitioned either in the filter energies coordinate space, or in any other reduced representation. The estimator can be then approximated by

$$\hat{S}_k = \sum_n \hat{S}_{k|n} \cdot P(\vec{S} \in n | \vec{S'}) \qquad (10)$$

where $\hat{S}_k$ is the MMSE estimator obtained for the n-th distribution component.

III. RECOGNITION EXPERIMENTS

The above algorithms were evaluated with SRI's DECIPHER continuous-speech, speaker-independent, HMM recognition system [8]. The system's acoustic front-end consists of performing 512-point FFT on 25.6 msec long speech frames, every 10 msec. The spectral power is summed in frequency bands corresponding to 25 overlapping Mel-scale filters, spanning a frequency range of 100-6400 Hz. A discrete cepstral transform is performed on the filter energies. The HMM is trained with discrete densities of four features: the truncated cepstral vector C_1-C_{12}, the DC component C_0, and their corresponding time derivatives (Delta). The vector quantization of the cepstral and delta-cepstral vectors use a variance-weighted euclidean distance metric. The recognition task was the 1000 word vocabulary of the DARPA-NIST "Resource management task" using a word-pair grammar (perplexity 60) [11].

The training was based on 3990 sentences of high quality speech, recorded at TI in a sound attenuated room with a close-talking microphone (designated by NIST as the Feb. '89 large training set). The testing material consisted of two types of noisy speech. The first was the DARPA-NIST "Resource management task" February 1989 test set (30 sentence from each of 10 talkers not in the training set), with additive white gaussian noise. The second consisted of recordings made at SRI, with both close talking and desktop microphones, in a noisy environment. Nine speakers participated in the SRI recordings, each speaking 30 sentences from the "Resource management task". The noise, predominantly generated by air conditioning, was estimated from a three seconds sample, recorded at the beginning of each speaker session. For these recordings the estimation algorithm was supplemented with equalization, to compensate for global differences between the SRI microphones and the one used for the training database. The equalization was particularly necessary for the desktop microphone, whose frequency response was very much dependent on the location of the speaker relative to the microphone. The equalization was performed by aligning each speaker's long term average spectrum with that obtained by averaging the spectrum over the whole training database.

Fig. 1 shows the recognition error rate obtained with no processing and with our best estimation algorithm (Eq. (10)), when trained on clean speech and tested with additive white noise. The SNR is defined here as the ratio between signal and noise average power, computed directly on the waveform. The performance without processing at 23 dB SNR is almost equal to that achieved with preprocessing at 10 dB SNR, suggesting that the estimation improves the effective SNR by 13 dB.

Fig. 2 compares the error rate achieved with several estimation algorithms, all tested on the TI recorded speech with additive white noise at 10 dB SNR. With the exception of the two rightmost charts, the training was performed on clean speech. The estimation algorithms are, from left to right: (1) no processing; (2) filter energy subtraction following Van Compernolle's method [6] where, whenever the noisy filter-energy is below the noise level, it is fixed to 50 dB below the highest observed energies for that filter; (3) MMSE of the logarithm of DFT magnitude, following the method of Porter and Boll [5]; (4) our MMSE estimation of filter log-energies (Eq. (3)); (5) our improved algorithm (Eq. (10)); (6) train on the TI recorded database with additive white gaussian noise at 10 dB SNR, without any processing; (7) train in noise at 10 dB SNR with our improved estimation algorithm.

Summarizing the results in fig. 2, the performance with the filter-bank MMSE is equivalent to that with log|DFT| MMSE, both achieving error rate which is approximately twice that obtained when the training is performed under exactly the same noise conditions as the testing. The improved filter-bank estimator reduces the error rate to only 50% above the training in noise. Finally, the estimation improves the performance even when the training and testing are done under exactly the same conditions.

Fig. 3 shows the error rate for the SRI recordings with the close-talking and desktop microphones. The average SNR, computed on the waveforms, was 31 dB for the close-talking and 13 dB for the desktop. The

speaker-average SNRs in individual filters, averaged over all filters, were 32 and 23 dB, respectively. Error rate is given with no processing, with our best estimation algorithm, and with both estimation and equalization. With both estimation and equalization, the error rate with the desktop microphone is only 40% higher than that with the close-talking microphone.

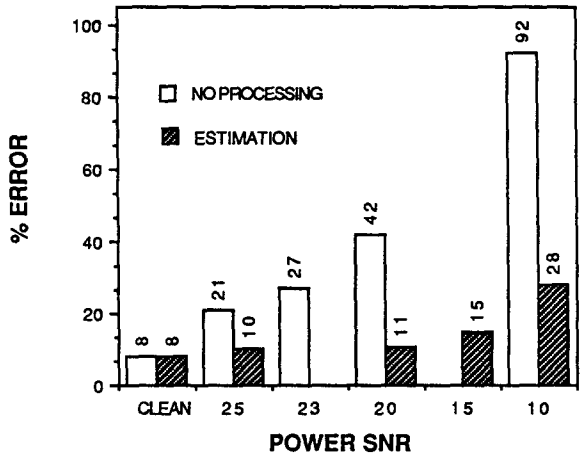

Figure 1.

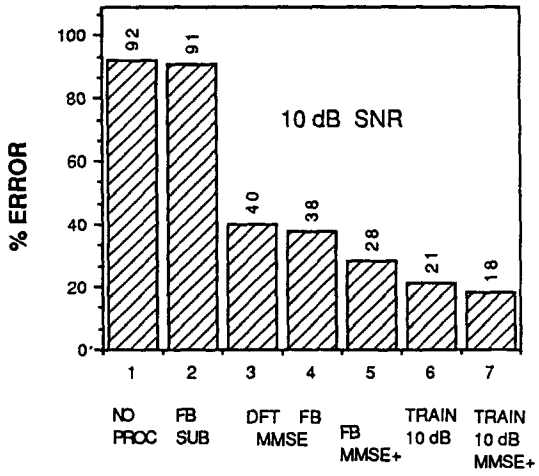

Figure 2.

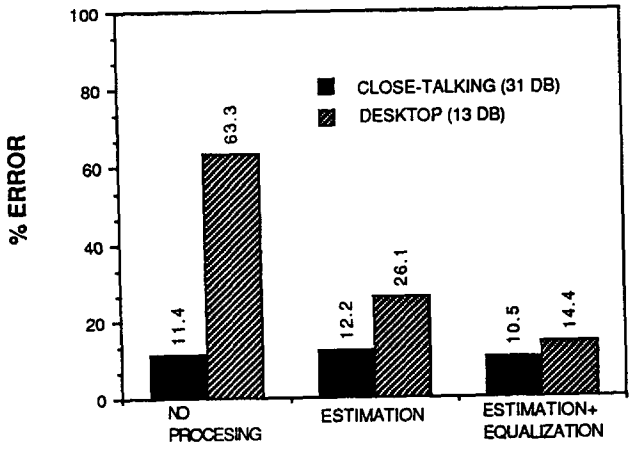

Figure 3.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grant IRI-8720403, and in part by SRI internal funding.

REFERENCES

1. S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. ASSP*, vol. 27, pp. 113-117, April 1979.

2. R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter," *IEEE Trans. ASSP*, vol. 28, pp. 137-140, April 1980.

3. Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. ASSP*, vol. 32, pp. 1109-1112, December 1984.

4. Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans.ASSP*, vol. 33, pp. 443-447, April 1985.

5. J. E. Porter and S. F. Boll, "Optimal Estimators for Spectral Restoration of Noisy Speech," *Proc. ICASSP*, vol. 2, pp. 18A2.1 - 2.4, 1984.

6. D. Van Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System," *Computer Speech and Language*, vol. 3, pp. 151-167, 1989.

7. D. Van Compernolle, "Spectral Estimation Using a Log-Distance Error Criterion Applied to Speech Recognition," *Proc. ICASSP*, vol. 1, pp. 258-261, 1989.

8. M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," *Proc. ICASSP*, vol. 1, pp. 672-699, 1989.

9. P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," Proc. ICASSP, vol. 1, pp. 651-654, 1988.

10. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, New York: John Wiley & Sons, Inc., 1966, p. 374.

11. W. M. Fisher, G. R. Doddington and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specification and Status," *Proc. DARPA Speech Recognition Workshop*, pp. 93-99, February 1986.