Recognition of Noisy Speech: Using Minimum-Mean Log-Spectral Distance Estimation

A. Erell and M. Weintraub
SRI International
333 Ravenswood Ave.
Menlo Park, CA 94025

Abstract

A model-based spectral estimation algorithm is derived that improves the robustness of speech recognition systems to additive noise. The algorithm is tailored for filter-bank-based systems, where the estimation should seek to minimize the distortion as measured by the recognizer's distance metric. This estimation criterion is approximated by minimizing the Euclidean distance between spectral log-energy vectors, which is equivalent to minimizing the nonweighted, nontruncated cepstral distance. Correlations between frequency channels are incorporated in the estimation by modeling the spectral distribution of speech as a mixture of components, each representing a different speech class, and assuming that spectral energies at different frequency channels are uncorrelated within each class. The algorithm was tested with SRI's continuous-speech, speaker-independent, hidden Markov model recognition system using the large-vocabulary NIST "Resource Management Task." When trained on a clean-speech database and tested with additive white Gaussian noise, the new algorithm has an error rate half of that with MMSE estimation of log spectral energies at individual frequency channels, and it achieves a level similar to that with the ideal condition of training and testing at constant SNR. The algorithm is also very efficient with additive environmental noise, recorded with a desktop microphone.

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. Several methods for handling this problem are in common use, among them supplementing 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.

The problem of designing a statistical estimator for speech recognition is that of defining an optimality criterion that will match the recognizer, and deriving an algorithm to compute the estimator based on this criterion. Defining the optimal criterion is easier for speech recognition than it is for speech enhancement for human listeners, since the signal processing is known in the former, but not in the latter. For a recognition system that is based on a distance metric, whether for template matching or vector quantization, a reasonable criterion would be to minimize the average distortion as measured by the distance metric. In practice, achieving this criterion may turn out not to be feasible, and the question is then to what extent the computationally feasible methods approximate the desired optimality criterion.

A basic difference between the cepstral distance criterion and the MMSE of single frequency channels (whether DFT coefficients or filter energies) is that the former implies a joint estimate of a feature vector, whereas the latter implies an independent estimation of scalar variables. Because the speech spectral energies at different frequencies are correlated, an independent estimate of individual channels results in a suboptimal estimation. To incorporate part of the correlations in the estimator, we modified our single-channel MMSE to be conditioned on the total energy in addition to the filter energy. This modification indeed improved performance significantly.

We here derive a more rigorous method of approximating the cepstral distance criterion. The optimality criterion is the minimization of the distortion as measured by the Euclidean distance between vectors of filter log energies. We name the algorithm *minimum-mean-log-spectral-distance* (MMLSD). The MMLSD is equivalent to minimizing the nonweighted, nontruncated cepstral distance rather than the weighted, truncated one used by the recognizer. The necessity for this compromise arises from the difficulty in modeling the statistics of additive noise in the transform domain, whereas a model can be constructed in the spectral domain [for details see Eq. (2): the approximation there will not work for the transformed vector].

The MMLSD estimator is first computed using a stationary model for the speech spectral probability distribution (PD). The PD of the filter log-energy vectors is assumed to comprise a mixture of classes, within which different filter energies are statistically independent. Several implementations of this model are considered,

including vector quantization and a maximum-likelihod fit to a mixture of Gaussian distributions.

II. Minimum-mean log-spectral distance estimation

The MMSE on the vector $\vec{S}$ of K filter log-energies yields the following *vector* estimator

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

where $\vec{S'}$ is the observed noisy vector, $P(\vec{S})$ is the clean speech log-spectral vector PD, and $P(\vec{S'} \mid \vec{S})$ is the conditional probability of the noisy log-spectral vector given the clean. This estimator is considerably more complex than the independent MMSE of single channels because it requires an integration of K-dimensional probability distributions. However, its computation can proceed using the following models for $P(\vec{S'} \mid \vec{S})$ and $P(\vec{S})$.

The conditioned probability $P(\vec{S'} \mid \vec{S})$ can be modeled simply as the product of the marginal probabilities,

$$P(\vec{S'} \mid \vec{S}) = \prod_{k=1}^{K} P(S'_k \mid S_k) \tag{2}$$

where $P(S'_k|S_k)$ is given in [1]. This factorization is a reasonable approximation because the noise is uncorrelated in the frequency domain and because, for additive noise, the value of a given noisy filter energy, S'_k, depends only on the clean energy S_k and on the noise level in that frequency. This model is obviously only an approximation for overlapping filters.

A similar factorization of $P(\vec{S})$ would lead to MMSE of individual frequency channels. However, such a factorization would be very inaccurate because the speech signal is highly correlated in the frequency domain. A more accurate model that partly incorporates the correlations between frequency channels is the following mixture model:

$$P(\vec{S}) = \sum_{n=1}^{N} C_n P_n(\vec{S}), \quad P_n(\vec{S}) = \prod_{k=1}^{K} P_n(S_k) \tag{3}$$

the idea being that the acoustic space can be divided into classes within which the correlation between different frequency channels is significantly smaller than in the space as a whole. An easily implemented parameterization would be to model the probabilities $P_n(S_k)$ as Gaussian with means μ_{nk} and standard deviations σ_{nk}. The classes can represent either mutually exclusive or overlapping regions of the acoustic space. The estimator is now given by

$$\hat{S}_k = \sum_{n=1}^{N} \hat{S}_k \mid n \cdot P(n \mid \vec{S'}) \tag{4}$$

where the first term is the n^{th} class-conditioned MMSE estimator, computed similarly to Eq. (2) with P(Sk) replaced by $P_n(S_k)$:

$$\hat{S}_k|n = \frac{1}{P(S'_k \mid n)} \int S_k\, P(S'_k \mid S_k)\, P_n(S_k)\, dS_k \tag{5a}$$

$$P(S'_k \mid n) = \int P(S'_k \mid S_k)\, P_n(S_k)\, dS_k \tag{5b}$$

and the second term is the *a posteriori* probability that the clean speech vector belonged to the n^{th} class, given by

$$P(n \mid \vec{S'}) = \frac{C_n P(\vec{S'}|n)}{\sum_{n=1}^{N} C_n P(\vec{S'}|n)} \tag{6a}$$

where

$$P(\vec{S'}|n) = \prod_{k=1}^{K} P(S'_k \mid n) \tag{6b}$$

Thus the estimator is a weighted sum of class-conditioned MMSE estimators.

III. Speech-recognition experiments

We evaluated the above algorithms with SRI's DECIPHER continuous-speech, speaker-independent, HMM recognition system [2]. The recognition task was the 1,000-word vocabulary of the DARPA-NIST "Resource management task" using a word-pair grammar with of perplexity 60 [3]. The training was based on 3,990 sentences of high-quality speech, recorded at Texas Instruments in a sound-attenuated room with a close-talking microphone (designated by NIST as the February 1989 large training set).

The testing material was from the DARPA-NIST "Resource Management Task" February 1989 test set [3] and consisted of 30 sentences from each of 10 talkers not in the training set, with two types of additive noise. The first is a computer-generated white Gaussian noise, added to the waveform at a global SNR of 10 dB. The SNR in individual frequency channels, averaged over all channels

and speakers, was 9 dB. The second is environmental noise recorded at SRI's speech laboratory with a desktop microphone. The environmental noise was quasi stationary, predominantly generated by air conditioning, and had most of its energy concentrated in the low frequencies. The noise was sampled and added digitally to the speech waveforms with global SNR of 0 dB; the SNR in individual frequency channels, averaged over all channels and speakers, was 12 dB.

The experiments in the environmental noise have been conducted both with and without tuning of the estimation algorithms to this particular noise. The tuning consisted of adjusting the degrees-of-freedom parameter in the chi-squared model, for the noise-filter energy, wide-band energy and total energy. Without tuning, the parameter values were those determined for white noise. A significant difference between the degrees of freedom for white noise and for environmental noise was found for the total-energy model: Because most of the environmental noise energy concentrated in the low frequencies, the number of degrees of freedom was very small compared to that with white noise. Only minor differences were found for the wide-band energies, and even smaller differences for the filter log energies.

Table 1 lists for reference the error rates with and without additive white Gaussian noise at 10-dB SNR, without any processing and with MMSE estimation. Table 2 lists error rates with white Gaussian noise, comparing the single-frame MMLSD algorithm with four mixture models, as a function of the number of classes N. With N=1, all the mixture models are identical to the MMSE estimator whose performance is given in Table 1. MMLSD-VQ and GM achieve the lowest error rates, with an insignificant edge to MMLSD-GM. The performance of both algorithms improves slowly but significantly when the number of classes N increases from 4 to 128. MMLSD-TE achieves error rates comparable to MMLSD-WB, and both algorithms reach a plateau in their performance level with N=4. MMLSD-TEP, with the total energy computed on the preemphasized waveform, does not perform as well as MMLSD-TE.

Summarizing the results, when training on clean speech and testing with white noise, the best MMLSD algorithm achieves the same error rate as training and testing in noise. In comparison, the error rate with MMSE is twice as high. Replacing the static mixture model by a dynamic Markov one makes no significant improvement. The error rates with environmental noise for the various algorithms are very similar to those with white noise, indicating that the algorithms are effective to a similar degree with the two types of noise.

IV. Discussion

A. Validity of the mixture model

The MMLSD estimator computed using the mixture model is much superior to the single-channel MMSE, indicating that the mixture model is successful in incorporating correlations between different frequency channels into the estimation. An interesting question, however, is to what extent the underlying assumption of the mixture model is correct: that is, is the statistical dependence between different frequency channels indeed small within a single mixture component. Measuring correlations between frequency channels with overlapping filters, we found that this assumption is incorrect. For example, with the vector quantization method (MMLSD-VQ) and a code book of size 32, the correlation between any pair of adjacent channels is of the order of 0.8, dropping to 0.4 for channels that are 3 filters apart and to 0.1 for channels that are 8 filters apart. The Gaussian mixtures model (MMLSD-GM) did not reduce the correlations: the maximum likelihood search converged on parameters that were very similar to the initial conditions derived from the vector quantization. The recognition accuracy obtained with MMLSD-GM is indeed identical to MMLSD-VQ.

Examining the MMLSD estimator in Eq. (4), we find that it is the *a posteriori* class probability that is erroneously estimated because of the invalid channel-independence asumption, Eq. (6b). The error in estimating this probability is magnified by the high number of channels: Small errors accumulate in the product Eq. (6b) of the assumedly independent marginal probabilities. In contrast to Eq. (6b), the output PD for the nonoverlapping wide bands is more accurate. With 3 bands and 32 classes the correlation between energies of different bands is approximately 0.15. Thus, although the overall MMLSD-WB estimator is not more accurate than MMLSD-VQ, the *a posteriori* class probability is more accurately estimated in MMLSD-WB than in MMLSD-VQ.

B. Total energy

The classification according to total energy, computed without preemphasis (MMLSD-TE), achieved excellent results with white noise but did not do as well as the other algorithms with the environmental noise. This result can be explained by the different SNRs in the two cases: whereas the total energy was 10 dB with the SNR white noise, it was 0 dB with the environmental noise. Because the degree to which the *a posteriori* class probability $P(n \mid E')$ peaks around the true class depends on the SNR in the total energy, it not surprising that MMLSD-TE was efficient for white but not for environmental noise.

A similar argument explains the advantage of MMLSD-TEP (where the total energy is defined on the

preemphasized waveform) over MMLSD-TE for the environmental noise, and the reverse for white noise: The average SNR on the preemphasized waveforms was 12 dB for the environmental noise and 3 dB for white noise. However, it seems that in no case is MMLSD-TEP as efficient as MMLSD-TE is with white noise.

C. Relation to adaptive prototypes

If one augments the MMLSD estimator with a detailed word-, or phoneme-based, continuous-density HMM, that model itself can be used for the speech recognition task. Instead of preprocessing the speech, optimal recognition would be achieved by simply replacing the clean speech output PDs by the PDs of the noisy speech, Eq. (6b). Another, computationally easier alternative is to adapt only the acoustic labeling in a semicontinuous HMM. Nadas et al. [4] used such an approach: their HMM was defined with semicontinuous output PDs, modeled in the spectral domain by tied mixtures of diagonal covariance Gaussians. The acoustic labeling was performed by choosing the most probable prototype given the signal. The same procedure was used in noise, modifying the output PDs to account for the noise. A similar procedure can be used with the model presented here: all that is required for acoustic labeling in noise is choosing n that maximizes $P(n \mid S')$, where the latter is given by Eq. (6). The difference between our model and that of Nadas et al. will then be only that they use the approximate MIXMAX model for $P(S'_k \mid n)$, whereas we will use the more accurate model in Eq. (5b).

The above approach would have an advantage over preprocessing by estimation if the HMM can indeed be designed with output PDs in the spectral domain and with diagonal covariance matrices. Unfortunately, it is currently believed that for speech recognition defining the PDs in the spectral domain is much inferior to the transform domain. It is for HMMs in the transform domain that the MMLSD preprocessing should be used.

V. Conclusions

We presented an estimation algorithm for noise robust speech recognition, MMLSD. The estimation is matched to the recognizer by seeking to minimize the average distortion as measured by a Euclidean distance between filter-bank log-energy vectors, approximating the weighted-cepstral distance used by the recognizer. The estimation is computed using a clean speech spectral probability distribution, estimated from a database, and a stationary, ARMA model for the noise.

The MMLSD is computed by modeling the speech-spectrum PD as a mixture of classes, within which different frequency channels are statistically independent. Although the model is only partially successful in describing speech data, the MMLSD algorithm proves to be much superior to the MMSE estimation of individual channels, even with a small number of classes. A highly efficient implementation of the mixture model is to represent the speech spectrum by a small number of energies in wide frequency bands (three in our implementation), quantizing this space of wide-band spectrum and identifying classes with code words. This method achieves performance that is almost comparable to that of a Gaussian-mixture model, at a much smaller computational load.

When trained on clean speech and tested with additive white noise at 10-dB SNR, the recognition acuracy with the MMLSD algorithm is comparable to that achieved with training the recognizer at the same constant 10-dB SNR. Since training and testing in constant SNR is an ideal situation, unlikely ever to be realized, this is a remarkable result. The algorithm is also highly efficient with a quasi-stationary environmntal noise, recorded with a desktop microphone, and requires almost no tuning to differences between this noise and the computer-generated white noise.

Acknowledgments

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

References

1. A. Erell and M. Weintraub, "Spectral estimation for noise robust speech recognition," DARPA Speech and Natural Language Workshop, October 1989.

2. M. Cohen, H. Murveit, P. Price, and M. Weintraub, "The DECIPHER speech recognition system," *Proc. ICASSP*,,1 (1990), S2.10.

3. P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," *Proc. ICASSP* 1, 651-654, 1988.

4. A. Nadas, D. Nahamoo, and M.A. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. on ASSP* 37, No. 10 (October 1989).

Algorithm and Noise Conditions	Percent Error
Train clean, test clean	8
Train clean, test in noise:	
No processing	92
MMSE	38
Train and test in noise, no processing	21

Table 1. Word error rate with and without MMSE estimation, for several noise conditions.

Model	Number of Classes			
	4	12	32	128
MMLSD-VQ	25.0	—	22.7	—
MMLSD-GM	24.7	—	21.9	21.0
MMLSD-WB (3 bands)	26.3	—	25.2	—
MMLSD-TE	25.1	25.3	—	—
MMLSD-TEP	—	34.3	—	—

Table 2. Word error rate with digital white noise at 10 dB SNR using a single-frame MMLSD estimation, as a function of the number of classes (mixture components) for the different mixture models.

Algorithm	Error Rate	
	Untuned	Tuned
No processing	84.6	—
MMSE	32.2	32.2
MMLSD-VQ (N=32)	18.5	18.5
MMLSD-WB (N=32)	20.4	19.7
MMLSD-TE (N=12)	32.4	27.5

Table 3. Word error rate with added noise recorded by a desktop microphone at 0 dB SNR; tuning refers to adjusting the noise-model parameters (number of degrees of freedom) from their values in white noise to their best values in the environmental noise.