RECONSTRUCTION OF DAMAGED SPECTROGRAPHIC FEATURES FOR ROBUST SPEECH RECOGNITION

Bhiksha Raj[1]*, Michael L. Seltzer*[2]*, and Richard M. Stern*[2]
1. Compaq Computer Corporation
Cambridge, MA 02142 USA
2. Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA

ABSTRACT

We present two missing-feature based algorithms that recover noise-corrupted regions of spectrographic representations of speech for noise-robust speech recognition. These algorithms modify the incoming feature vector without any changes to the speech recognition system, in contrast to previously-described approaches. The first approach clusters the feature vectors representing clean speech. Missing data are recovered by estimating the spectral cluster in each analysis frame based on the uncorrupted feature values. The second approach uses MAP procedures to estimate the values of missing data elements based on their correlations with the features that are present. Both methods take into account bounds on the clean spectrogram implied by the noisy spectrogram. Large improvements in recognition accuracy are observed when these methods are used on speech corrupted by non-stationary noise when the locations of the corrupt regions of the spectrogram are known. We also present a new method of estimating the locations of corrupt regions in spectrograms that treats the problem of identifying these regions as one of Bayesian classification. This method, when used along with the best method to reconstruct them, results in recognition accuracies comparable with the best previous data compensation algorithm on speech corrupted by white noise. It also provides significant improvement on speech corrupted by music when the global SNR of the corrupted signal is known *a priori*.

1. INTRODUCTION

Speech recognition accuracy degrades when the speech to be recognized is corrupted by noise (*e.g.* [1]), especially when the recognition system itself has been trained on clean speech. Several methods have been proposed in the literature to compensate for the effect of noise on recognition systems (*e.g.* [1]). However, almost all of them assume, either explicitly or implicitly, that the noise that is corrupting the speech is stationary, and therefore fail on speech that has been corrupted by non-stationary signals such as music.

The missing-feature approach appears to be particularly promising in its ability to compensate for non-stationary noise. This approach takes advantage of the fact that, since the signal energy of speech and noise are different in different frequency bands, noise affects different regions of a spectrographic representation of speech differently. As a result, spectrograms of noise-corrupted speech exhibit regions of high SNR, and other regions with relatively low SNR. Missing feature methods mark the low-SNR regions as missing (effectively erasing them from the spectrogram) and make explicit use of the regions of high SNR in the corrupted speech to compensate for the missing regions. The most comprehensive work using this approach has been reported by researchers at the University of Sheffield [2], although other research groups are active in this field as well [3].

The methods described in [2,3] exploit *a priori* knowledge of the statistical representations of clean (uncorrupted) speech as modelled by a speech recognition system. Mean-imputation methods [2,3] find MAP estimates of corrupted frequency bands, utilizing the statistics of clean speech. Marginalization-based methods [2,3] attempt to ignore the contribution of noise-corrupted bands completely. Both of these methods constrain the recognizer to work in the log-spectral domain, where the missing regions are identified. This is a serious drawback because, although missing-feature approaches are highly effective, recognition accuracy using log spectra is, as a rule, much worse than that obtained using other features such as cepstral coefficients.

In this paper we present two methods that reconstruct the damaged regions of the spectrogram as a preprocessing step, prior to recognition: cluster-based and correlation-based reconstruction. The reconstructed spectrogram can then be transformed to the feature set of choice, such as cepstra, and used with a standard recognizer. This has the combined advantages of permitting different kinds of recognizers to be used, as well as permitting the use of information or modeling structures that are not explicitly handled by the recognizer, such as the temporal correlations between components of the log-spectral vectors.

The algorithms presented in this paper use simple statistical characterizations of the distributions of log spectra of clean speech in order to reconstruct the noise-corrupted regions of spectrograms. We incorporate bounding constraints obtained from noisy speech to refine the estimation and evaluate them on speech corrupted by different noises. Experimental results indicate that these methods can be highly effective when the corrupted regions of the spectrogram are identified perfectly.

A fully-automated missing-feature based approach also needs to identify the corrupted regions of the spectrograms automatically. In this paper we also present a new method for identifying corrupt regions of the spectrogram that treats the problem of identifying corrupt regions as one of Bayesian classification. Experiments show that this method, when used in conjunction with the missing-feature methods presented in this paper, results in significant improvement in recognition accuracies when speech is corrupted by white noise. It is also effective on speech corrupted by music when the SNR of the corrupted signal is known *a priori*.

In Section 2 we outline the algorithms used to estimate missing regions of spectrograms. In Section 3 we describe the imposition of constraints on the estimation process. In Section 4 we describe classifier-based identification of damaged portions of the spectrogram. In Section 5 we present experiments to evaluate the reconstruction methods described in this paper. Finally, in Section 6 we present our conclusions.

2. RECONSTRUCTION OF MISSING REGIONS OF SPECTROGRAMS

A frame-based log-spectral representation of noise-degraded speech can be thought of as a spectrogram-like representation. Some regions of such a representation would be more corrupted by the noise than others. In this work we characterize the less noise-corrupted regions of the representation as “present”, or “reliable”, and the more corrupted regions as either “missing” or “unreliable”. The lower the SNR, the larger the fraction of elements that are missing. The goal of this work to reconstruct the missing/unreliable regions of the featural display from the information that is present/reliable, using whatever information is available.

We describe two techniques to reconstruct the missing portions of a corrupted spectrogram - cluster based reconstruction, and time-covariance based reconstruction [4]. In both techniques, we obtain *a priori* information about the structure of speech spectrograms from a training corpus of uncorrupted speech and utilize this information to arrive at an informed guess for the missing components of the corrupted speech log-spectral vectors. The two techniques differ in the kind of *a priori* information obtained and how it is applied to the reconstruction process.

2.1. Cluster Based Reconstruction

For cluster-based reconstruction, the *a priori* information about the speech signal is obtained by grouping all the log-spectral vectors from an uncorrupted training database into a number of clusters and finding the statistical parameters of each cluster. Clusters are assumed to have Gaussian distributions, and clustering is accomplished using conventional EM techniques [5]. The statistical properties of each of the clusters are the mean, the covariance, and the prior probability of the cluster.

To compensate noisy speech, the algorithm attempts to identify the cluster to which each log-spectral vector of the noise-corrupted speech belongs. This cluster is identified as the one with the greatest *a posteriori* probability of having generated the noise-corrupted vector, where the *a posteriori* probabilities of clusters are computed based solely on the values of the components of the log-spectral vectors that are present.

Once the cluster is identified, the covariance and mean of the vectors belonging to that cluster are used to obtain MAP estimates of the missing components of the vector.

2.2. Correlation-Based Reconstruction

In correlation-based reconstruction the sequence of log-spectral vectors of a speech utterance are assumed to be samples of a stationary Gaussian random process. The *a priori* information about the clean speech signal is represented by the statistical parameters of this random process: its mean and the covariances. These parameters are estimated from an uncorrupted training database.

To compensate corrupted speech where some components of the log-spectral vectors are missing, a vector X_t is formed that consists of all the elements that are missing in any log-spectral vector Y_t. A second vector, N_t, is formed of all the elements that are present anywhere in the spectrogram, that have a normalized correlation of at least 0.5 with at least one of the elements of the vector X_t. The value of X_t is now obtained as an MAP estimate conditioned on the vector N_t, as follows:

$$\hat{X}_t = M_t + R_{n,n} C_{n,n}^{-1} N_t \tag{1}$$

where $\hat{X}_t$ is the estimate for X_t, M_t is the mean of the distribution of X_t, $R_{n,n}$ is the covariance of X_t and N_t, and $C_{n,n}$ is the autocovariance matrix of the elements in the vector N_t. We refer to this method as *correlation-based reconstruction,* since the method explicitly makes use of both temporal and spectral correlations of the elements in the vector sequence.

3. ESTIMATION WITH BOUNDS

The noisy regions of the spectrograms of noisy speech are not completely devoid of information. If we assume that the corrupting noise is uncorrelated to the clean speech signal we get:

$$e_{ns} = e_n + e_s \tag{2}$$

where e_{ns} is the energy of the noisy signal, e_n is the noise energy, and e_s is the energy of the clean speech. Clearly e_n cannot be greater than e_{ns}. Hence e_{ns} gives us an upper bound on our estimate of e_s.

Conventional missing-feature techniques use this bounding information to constrain the distributions of the classes that the recognizer considers by setting the probability of all values that are greater than the bound to zero [2]. In this paper this bounding information is used to constrain both cluster-based reconstruction and correlation-based reconstruction. In cluster-based reconstruction, as in conventional missing-feature methods, the energy in the noisy spectrogram is used to constrain the distribution of the clusters when computing the *a posteriori* probabilities of clusters, thereby limiting the clusters that are considered in the reconstruction. Once the appropriate cluster is identified, the bounds are further used to constrain the MAP estimates of the missing components. In the case of correlation-based reconstruction the bounds constrain the MAP estimation procedure that is used to estimate the missing regions.

4. CLASSIFIER-BASED ESTIMATION OF SPECTROGRAPHIC MASKS

The performance of missing-feature methods is critically dependent on the accuracy with spectrographic masks - the tags which identify elements of the spectrogram as corrupt or reliable - are estimated. Traditionally, spectrographic masks are estimated from estimates of the local SNR of each of the elements in the noise spectrum The local SNR is, in turn, estimated from running estimates of the noise spectrum [6]. The accuracy of the estimated spectrographic masks is dependent on the accuracy with which the SNR is estimated.

Spectrographic masks are essentially tags that separate elements of the spectrogram into two classes – the class of reliable elements, and the class of corrupt ones. Each element of the spectrogram belongs to one of these two classes. In this paper we introduce *classifier-based estimation of spectrographic masks*, where we therefore treat the problem of estimating spectrographic masks as one of Bayesian classification.

Each element of the spectrogram is represented by a vector constructed from the value of the element itself and the differences between neighboring elements along each of four axes (left-right, up-down, and the two diagonal axes). Gaussian mixture distributions are trained for the each of the two classes using vectors derived from spectrograms of an artificially corrupted corpus of speech for which the true spectrographic masks are available. A separate classifier is trained for each of the frequency bands in the spectrogram. The *a priori* probability of the classes is empirically chosen to give optimal performance over a variety of SNRs.

Spectrographic masks for noisy speech spectrograms are estimated by classifying the vector representing each of the spectrographic elements in every frequency band using the classifier for that frequency band.

5. EXPERIMENTAL RESULTS

We evaluated the methods described above and others using the DARPA Resource Management (RM1) database, and compared them with the best conventional method, marginalization [2]. The log-spectral representation was developed from the outputs of twenty standard mel-scaled filters. Clusters and their statistics were obtained from the log-spectral representations of the training set of utterances for the cluster-based methods. For the correlation based method the means and covariances needed were also computed from the training corpus. For purposes of evaluation we corrupted speech with white noise and with segments of music from the “Marketplace” radio program. The local SNR was known to the system. The training set consisted of 2880 utterances from the training set of RM1. The test set consisted of 1600 RM1 evaluation utterances. The SPHINX-3 speech recognition system was used to train 1 gaussian/state HMMs. No delta or double delta features were used. For all experiments not involving marginalization, mean normalization of the features was performed. No mean normalization was performed in experiments involving marginalization since mean normalization degrades its performance badly.

Two sets of experiments were run. In the first set of experiments the performance of spectrogram reconstruction methods was evaluated, and compared with that of marginalization, when the spectrographic masks of noisy speech spectrograms were known *a priori*. In the second set of experiments the performance of spectrogram reconstruction methods with estimated spectrographic masks was evaluated.

5.1. Experiments with perfect knowledge of spectrographic masks

In the experiments reported in this section we evaluate the performance of the spectrogram reconstruction methods described in this paper, as well as that of marginalization, under ideal conditions, *i.e.* when the true SNR of the spectrographic elements is available for generating spectrographic masks. For the spectrogram reconstruction methods described in this paper, spectrographic masks were generated by marking all regions in the spectrograms with SNR lower than -5 dB as missing or unreliable. For marginalization, a threshold of 15 dB was used to generate spectrographic masks using empirically-derived thresholds. Three sets of experiments were run. In the first set the effect of incorporating the upper bound as described in Section 3 was evaluated. In the second set the effect of spectral subtraction [7] on missing feature methods was evaluated. Finally, the performance of the proposed methods on cepstra based recognition was evaluated on speech corrupted with both white noise and music.

Bounding the estimates: We first consider the effect of bounding on the performance of cluster-based and correlation-based reconstruction methods, and marginalization. Figure 1a shows the performance of these algorithms when they are not bounded by the noisy spectrogram. Figure 1b shows the performance when the bounded estimation procedures described in Section 3 are used. The performance of all missing-feature reconstruction algorithms is clearly greatly improved by the application of bounds. The greatest improvement is seen on cluster-based reconstruction, which is completely ineffective when no bounds are applied, but becomes highly effective when bounds are used. Overall, we note that conventional marginalization outperforms the methods proposed in this paper at most SNRs.

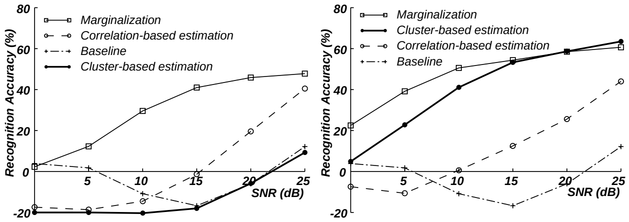

Figure 1. a) Recognition accuracy obtained with missing feature methods without bounding. b) Recognition accuracy when the energy in the noisy speech spectrogram is used as an upper bound.

Effect of spectral subtraction on missing-feature reconstruction: Figure 2a compares recognition accuracy obtained with cluster-based and correlation-based methods, marginalization, and spectral subtraction [7] on speech that has been corrupted by white noise. We observe that the accuracy obtained with spectrogram reconstruction methods proposed in this paper is only slightly greater than that obtained by spectral subtraction in isolation. The reason for this is that spectrogram reconstruction methods assume that the reliable (present) regions of the spectrogram are uncorrupted. However, this is not really true: the SNR in these regions can be as low as -5 dB. One solution to this problem is to attempt to reduce the noise level even in those portions of the spectrogram that are marked “present”, for example, by applying spectral subtraction to these regions. Figure 2b shows the recognition accuracies obtained when spectral subtraction is combined with the missing feature methods. As can be seen, the use of spectral subtraction before missing-feature reconstruction substantially improves accuracy. Note that since the SNR threshold used to generate spectrographic masks for marginalization is much higher than that used for the reconstruction methods, spectral subtraction does not affect its performance significantly. As a result, the recognition accuracy obtained with cluster-based estimation is now comparable with, or better than that obtained with marginalization at most SNRs.

Recognition using cepstra: As noted earlier, while conventional missing feature methods require that recognition be performed

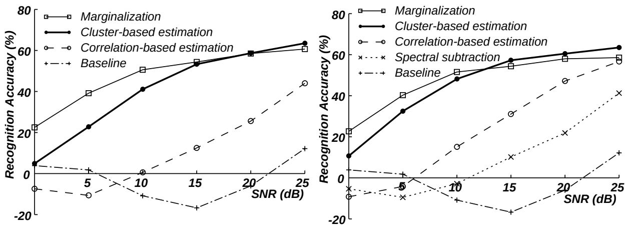

Figure 2. a) Recognition accuracy of missing feature methods when no spectral subtraction is employed b) Recognition accuracy when they are combined with spectral subtraction.

using log-spectral features, the missing-feature methods proposed in this paper result in reconstructed spectrograms from which cepstra can be computed, and cepstra based recognition performed. Figure 3a shows the recognition accuracy obtained on speech corrupted with white noise when recognition is performed using cepstra derived from reconstructed spectrograms. Comparison with Figure 2 shows that the performance obtained with cepstra derived using the methods proposed in this paper are much better than those obtained with marginalization and log-spectra based recognition. Figure 3b shows the performance of the proposed methods on speech corrupted at various SNRs by segments of music from the “Marketplace” news program. We note that spectral subtraction fails when speech is corrupted by music, even though it is effective in the presence of broadband noise. However, both correlation-based and cluster-based reconstruction result in large improvements in accuracy.

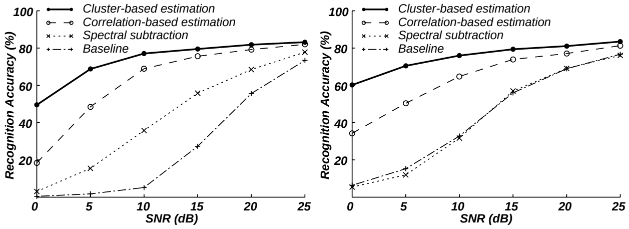

Figure 3. a) Performance of missing-feature methods on speech corrupted by white noise, when recognition is performed using cepstra. b) Performance of missing feature methods on speech corrupted by music, when recognition is performed using cepstra.

5.2. Experiments with estimated spectrographic masks

Figure 4 shows recognition accuracies obtained with cepstra derived from spectrograms reconstructed by cluster-based and correlation-based methods, using spectrographic masks that have been estimated by classifier-based estimation. This represents the true performance of missing feature based methods. We note that the reconstruction methods, when used with the estimated masks, result in large improvement in recognition accuracy on speech corrupted by white noise. On speech corrupted by music, classifier-based estimation of masks is ineffective. However, if the SNR of the music-corrupted speech is known *a priori*, (*i.e.* the classifier is trained for that specific SNR) classifier-based mask estimation results in significant improvements in recognition accuracy at all SNRs. This is the first time that improvements have been observed on speech corrupted by music.

6. SUMMARY AND CONCLUSIONS

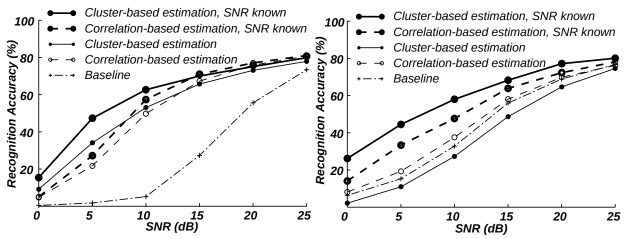

Figure 4. a) Recognition accuracy on speech corrupted by white noise, using estimated spectrographic masks. b) Recognition accuracy on speech corrupted by music using estimated masks.

In this paper we describe the performance of two new methods, cluster-based reconstruction and correlation-based reconstruction, that recover missing or unreliable feature information from log-spectral representations of noisy speech. These algorithms use simple characterizations of the statistical properties of log-spectral vectors to reconstruct noise corrupted regions of spectrograms. We observe better reconstruction, and hence better recognition accuracy if (1) reconstruction is constrained by the implicit bounds imposed on the estimate by noisy spectrograms, and (2) the reconstruction procedure is preceded by spectral subtraction. The proposed methods outperform conventional missing feature methods when recognition is performed using cepstra. We also present a new classifier-based method for identifying noisy locations in spectrograms. We demonstrate that the missing-feature algorithms are highly effective in compensating for stationary noises. Further, they are effective at compensating for non-stationary noises as well when the global SNR of the corrupted signal is known *a priori*.

ACKNOWLEDGMENTS

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. Moreno P. (1996) Speech Recognition in Noisy Environments, Ph. D. Dissertation, ECE Department, CMU, May 1996
2. Cooke, M.P., Morris, A. and Green, P. D (1996) “Recognizing Occluded Speech”, ESCA Tutorial and Workshop on Auditory Basis of Speech Perception, Keele University, July 15-19 1996
3. Lippman, R. P. (1997) “Using Missing Feature Theory to Actively Select Features for Robust Speech Recognition with Interruptions, Filtering and Noise”, Proc. Eurospeech 1997
4. Raj, B., Singh R., and Stern, R.M. (1998) “Inference of missing spectrographic features for robust speech recognition”, Proc. ICSLP 1998
5. A. P. Dempster, N. Laird, and D. B. Rubin, ``Maximum Likelihood from incomplete data via the EM algorithm," Journal of the Royal Statistical Society, 1977, Vol. B39
6. Hirsch, H.G., Ehrlicher, C. (1995), “Noise estimation techniques for Robust Speech Recognition”, Proc. IEEE Conf. on Acoustics, Speech and Signal Processing 1995
7. Boll, S.F. (1979), “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Transactions on Acoustics, Speech and Signal Processing, April, 1979