INFERENCE OF MISSING SPECTROGRAPHIC FEATURES FOR ROBUST SPEECH RECOGNITION

Bhiksha Raj, Rita Singh, and Richard M. Stern
Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA

ABSTRACT

Two types of algorithms are introduced that recover missing time-frequency regions of log-spectral representations of speech. These compensation 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 log-spectral vectors representing clean speech. Missing data are recovered by estimating the spectral cluster in each analysis frame on the basis of the feature values that are present. The second approach uses MAP procedures to estimate the values of missing data elements based on their correlation with the features that are present. Greatest recognition accuracy was obtained using the correlation-based approach, presumably because of its ability to exploit the temporal as well as spectral structure of speech. The recognition accuracy provided by these algorithms approaches but does not exceed that obtained by traditional marginalization. Nevertheless, it is believed that these algorithms provide greater computational efficiency and enable greater flexibility in recognition system structure.

1. INTRODUCTION

Automatic speech recognition systems perform poorly 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 reduce the damaging effects of noise on the performance of recognition systems (*e.g.* [1,4]) However, almost all of them make the assumption that the noise that is corrupting the speech in stationary. This is not necessarily a realistic expectation.

In addition, at any given instant of time, the signal energy in different frequency bands of a speech signal is different, so the degree of corruption, as measured by the signal-to-noise ratio (SNR), is different in each frequency band. Thus, the corrupted speech signal exhibits local regions (or “islands”) in the time-frequency plane, of relatively high SNR, as well as other islands with low SNR. Most standard methods of noise compensation do not take explicit advantage of this fact.

An alternative approach would be to make explicit use of the regions of high SNR in the corrupted speech to compensate for the islands of low SNR. The most comprehensive work using this approach has been reported by researchers at the University of Sheffield (*e.g.* [2]) and has also been described in [3].

An important disadvantage of the methods described in [2,3] is that they depend on the statistical representations of speech that are used by the speech recognizers as the *a priori* distributions of clean speech vectors. Mean-imputation based methods [2,3] find MAP estimates of corrupted frequency bands, utilizing the statistics of clean speech. Marginalization-based methods [2,3], on the other hand, attempt to ignore the contribution of noise-corrupted bands completely. Both of these methods are dependent on analytic statistical characterizations of the effects of the degradation on the internal statistical model used by the recognizer to represent speech.

The algorithms presented in this paper attempt to compensate for the effects of time-varying and transient noise by modification of the incoming features, rather than by modification of the speech recognizer to re-estimate or selectively ignore missing log-spectral bands. 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.

In this paper we present a series of methods that perform compensation by modifying a frame-based mel-frequency log-spectral representation of the incoming speech signal. In Section 2 we describe a series of algorithms which cluster the spectral profiles of clean speech and then attempt to estimate the cluster to which each frame of incoming noise-corrupted speech belongs, based on islands of reliability in the representation. In Section 3 we describe a set of simpler algorithms which estimate missing regions on the log-spectral representation based on observed *a priori* covariances of features in the representation across frequency and time. In Section 4 we present our major results and observations, and we conclude the paper in Section 5 with a series of suggestions of future work.

2. CLUSTER-BASED INFERENCE

In this section we discuss algorithms which cluster the log-spectral vectors of clean speech into a codebook using conventional EM methods. To compensate noisy speech, the algorithm attempts to identify the cluster to which each log-spectral vector of noise-corrupted speech belongs. The covariance and mean of the vectors belonging to that cluster are then used to obtain MAP estimates of the corrupted portion of the vector, conditioned on the uncorrupted portions.

The frame-based log-spectral representation of noise-degraded speech can be thought of as a spectrogram-like representation which some regions more corrupted than the others. In this work we characterize the less noise-corrupted regions of the representation as “present” and the more corrupted regions as either "dropped" or "missing". The fraction of elements of the sequence of log-spectral vectors that are missing is inversely related to the SNR. The goal of this work to reconstruct the missing regions of the featural display from the information that is present, using whatever information is available.

For our cluster-based schemes, 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 various statistical relations between the vectors belonging to each cluster. Clustering is accomplished using conventional EM, assuming that vectors in each of the clusters are distributed according to a Gaussian distribution. The statistical properties of each of the clusters are the mean, the covariance, and the prior probability of the cluster.

In the following subsections we describe several methods of finding the cluster membership of "damaged" feature vectors with partially missing data.

2.1. Cluster Identification Based on Marginalization

We used marginalization methods to identify the cluster that is most likely in the statistical sense to have generated the uncorrupted features of a damaged vector, while ignoring the contribution of the missing components. This method closely resembles the marginalization procedure described by the Sheffield group in [2], except that our work modifies only the incoming data stream, without modifying the statistical representation of speech in a recognition system.

Cluster identification based on marginalization becomes increasingly erroneous as the degree of damage increases (as measured by the fraction of the components of the vector that are missing) because the number of elements that are available to estimate the cluster identity diminishes. In the extreme case where an entire vector is damaged, there is no way of guessing its cluster identity at all. In such cases we arbitrarily select for the totally-corrupted frames the estimated cluster for the closest vector that is not fully damaged.

2.2. Interpolation along the Frequency Axis

We can generally avoid the problem of having to ignore missing elements in a vector entirely by finding the closest undamaged elements on either side of missing elements in a vector and estimating the missing elements by linear interpolation between the undamaged neighbors that surround them. Since our vectors are log-spectral vectors, we refer to this procedure as *interpolation along the frequency axis*.

In practice we find that obtaining preliminary estimates for missing elements by interpolating along the frequency axis does not increase the probability of in determining the correct cluster identity. As in the case of marginalization-based cluster identification, it is not possible to estimate the cluster identity if the entire vector is damaged since there are no neighbors available for interpolation.

2.3. Interpolation along the Time Axis

In a fashion similar to frequency interpolation, a preliminary estimate for missing elements in a vector can be made by interpolating between the closest undamaged elements of the same frequency from adjacent or nearby time frames, in both directions. We refer to this procedure as *interpolation across the time axis*. This procedure has the advantage that an estimate for the elements of a vector is possible even when entire log-spectral vectors are damaged.

Recognition accuracy can be further improved by iterating the process of estimating missing and/or damaged data via interpolation. Specifically, the interpolation along the time axis, cluster identification, and MAP estimation of missing elements is repeated twice. Since there are no more missing elements during the second pass (because all elements that were originally labelled are now missing), damaged components of the vector are estimated by simply averaging the corresponding frequency components of the vectors immediately preceding and following the vector in question. Further iterations are possible as well.

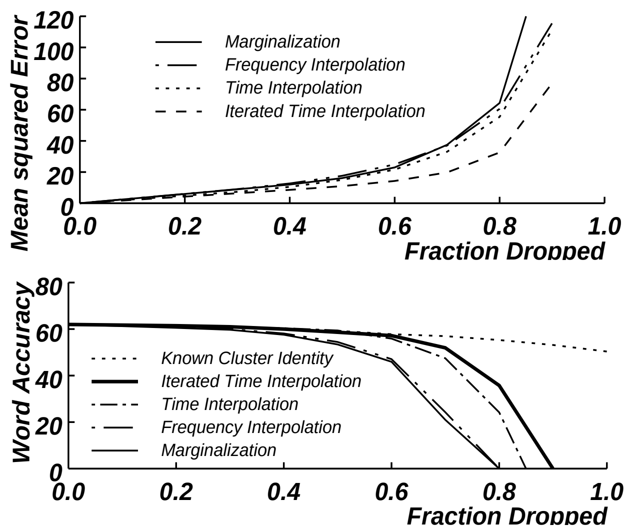

Figure 1. Dependence of mean squared error in cluster identification (upper panel) and recognition accuracy (lower panel) as a function of drop rate.

2.4. Experimental Results

We evaluated the methods described above and others using the DARPA Resource Management (RM1) database. 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 purposes of evaluation we randomly dropped elements of the log-spectral vectors of the test set, leaving untouched the remaining elements. The fraction of elements dropped is referred to as the "drop rate". In all experiments the locations of dropped elements were known to the system.

The upper panel of Figure 1 shows the mean-square difference (MSE) between the correct cluster location and the estimated cluster location according to the various methods described. We note that the use of frequency interpolation does not decrease the MSE much beyond the MSE obtained with marginalization. Interpolating along time does reduce the MSE substantially, and iterated interpolation along time provides the lowest MSE.

The lower panel of Figure 1 describes the results of speech recognition experiments using the same compensation techniques. The SPHINX-3 speech recognition system was used, with one Gaussian per state HMMs trained using the 2880 utterances from the speaker-independent training set of RM1. The test set consisted of the 1600 RM1 evaluation utterances. Word accuracies obtained when no compensation was performed degraded very quickly to 0 by the time only 10% of the elements were corrupted, and are therefore not shown. As can be seen in the lower panel of Figure 1, recognition accuracy obtained with cluster-based inference follows the patterns of MSE describe above. Specifically, frequency interpolation provides only slight improvement beyond direct marginalization. Interpolation across time provides substantially greater accuracy, and iterated temporal interpolation provides the best accuracy of all. One possible reason for the limited success of frequency interpolation is that this interpolation does not add much new information that is not already represented by the shapes of the spectral clusters. Temporal interpolation, on the other hand, exploits temporal continuity constraints, adding information that is complementary to the spectral clusters.

The upper dotted curve describes the recognition accuracy obtained when the correct cluster identity is used by the system for recognition; this curve represents the theoretical upper limit of accuracy to be expected from cluster-based inference. Comparison of this curve with the results of iterated temporal interpolation indicates that substantially better recognition accuracy could be obtained at the highest drop rates if we were able to improve the mechanism for identifying the cluster membership of the damaged vectors.

The fact that interpolation along time as a preliminary step to cluster identification results in an improvement in recognition accuracy leads us to believe that temporal correlations between vectors are a feature that can be exploited for reconstructing the damaged portions of the vector. This possibility is explored by the algorithm defined in the next section.

3. CORRELATION-BASED INFERENCE

The algorithms presented in this section differ from those described in Section 2 in that we form inferences on the basis of *a priori* statistical correlations of the elements of the log-spectral vectors, without regard to spectral clusters or other structural attributes. In these algorithms we assume that the sequence of feature vectors are samples of a stationary Gaussian random process, characterized by their means and covariances. These statistics can be used in conjunction with the elements that are present in the sequence of log-spectral vectors to estimate the missing elements in the sequence. We refer to this method as *estimation based on temporal correlation*, since the method explicitly makes use of temporal correlations of the elements in the vector sequence, along with spectral correlations.

3.1. Estimation based on Temporal Correlation

In this method we estimate a particular data element on the basis of *a priori* correlations between it and the most highly correlated elements that remain present in a damaged speech vector sequence. The *a priori* mean and the covariances from which the correlations are derived from an uncorrupted training database. To compensate corrupted speech, a vector X_t is formed that consists of all the elements that are missing in any log-spectral vector Y_t. In addition, a second vector, N_t, is formed that consists of all the elements that are present that have a normalized correlation of at least 0.5 with at least one of the elements of the vector X_t. We refer to the elements of N_t as elements that are in the "neighborhood" of a given missing element.

The value of X_t is now obtained as an MAP estimate conditioned on the vector N_t, as follows:

$$\hat{X}_t = M_t + R_{x,n} C_{n,n}^{-1} N_t \qquad (1)$$

where $\hat{X}_t$ is the estimate for X_t, M_t is the mean of the distribution of X_t, $R_{x,n}$ 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.

In principle, all the elements in the neighborhood of the vector being considered could be used in the estimation. However, we observed in practice that the use of just the ten to twelve most highly correlated elements was sufficient to obtain best accuracy.

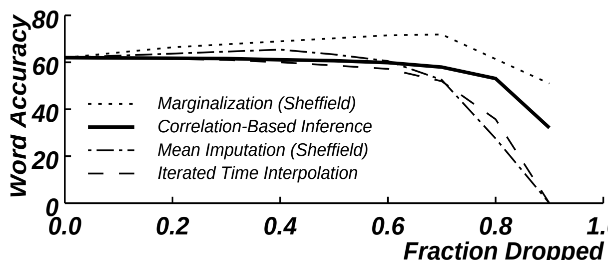

Figure 2. Comparison of recognition accuracy using correlation-based inference with our best cluster-based inference algorithm, along with two algorithms from the University of Sheffield [2]. (See text.)

3.2. Experimental Results

The recognition accuracy obtained using correlation-based inference is plotted in Figure 2 as a function of drop rate, using the same experimental setup as was described in Section 2.4. For comparison we also include in Figure 2 corresponding results for our best cluster-based compensation scheme, iterated interpolation along the time axis, as well as results obtained using local implementations of two of the best methods described by the Sheffield group, *mean imputation*, and *marginalization.* As noted above, the two Sheffield algorithms modify the speech recognizer in addition to the incoming feature vector, while out algorithms modify the incoming features only. Specifically, mean imputation uses the state distributions of the HMMs to reconstruct missing elements, while marginalization simply ignores the contributions of missing regions to the recognition process.

We note that our best cluster-based method, iterated interpolation along the time axis, provides recognition accuracy that is comparable to Sheffield's mean-imputation method. While estimation based on temporal covariance works considerably better than iterated interpolation along the time axis, it does not quite achieve the recognition accuracy obtained using Sheffield's marginalization algorithm.

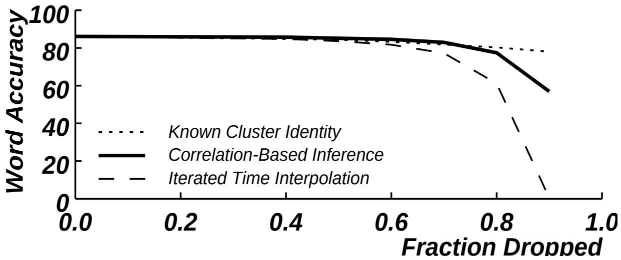

Figure 3. Recognition accuracy vs. drop rate for various compensation schemes using a combination of cepstra, delta cepstra, and double delta cepstra as the feature.

3.3. Feature Selection and Computation

While the data shown in Figure 2 show that traditional marginalization produces the greatest recognition accuracy, the methods introduced in Sections 2 and 3 can provide significant computational advantages. Most speech recognition systems use cepstral rather than log-spectral features. For example, Figure 3 describes the recognition accuracy obtained using a feature set consisting of traditional cepstra, delta cepstra and double delta cepstra. Compensation was performed in the log-spectral domain, with our most successful cluster-based and correlation-based algorithms, iterated temporal interpolation and temporal correlation. We note that absolute recognition accuracy is substantially better than was obtained using log-spectral features, and that the compensation algorithms provide improvements in accuracy that are comparable to the results described in Sections 2 and 3.

Nevertheless, compensation for missing features is accomplished in the log-spectral rather than the cepstral domain. If a recognition system uses a statistical representation of cepstral feature, the parameters of the state distributions for the speech recognition system must be transformed back to the log-spectral domain before either mean imputation or marginalization can be performed. If we assume that differenced cepstra and double-differenced cepstra are also included in the feature set, we estimate that traditional marginalization would require in each analysis frame an average of 10 inversions with about $90N^3$ multiplications to obtain these parameters, where N is the size of the log-spectral vector. (This estimate also assumes no more than 10 states, each with a 1 Gaussian per state distribution are being estimated at each instant.) Mean imputation would require an additional matrix inversion for the MAP estimation, for each Gaussian being considered.

The cluster-based methods (Section 2) require the inversion of only one matrix of order no greater than $(N-1) \times (N-1)$ needing only $(N-1)^3$ multiplications. The temporal-correlation method (Section 3) requires the inversion of one matrix of order no greater than $3N \times 3N$, or no more than $(3N)^3$ multiplications, assuming that 3 neighbors are used in the estimation per missing element, on average.

Another advantage of the procedures described in this paper lies in the actual estimation of the difference parameters. Difference cepstra are a linearly transformed version of differenced log spectra. Hence, the loss of either of the two log-spectral vectors that underlie a particular differenced log spectral frame would cause the corresponding differenced cepstral coefficient to be considered missing. As a result, the fraction of differenced log cepstral elements that are missing can, in the worst case, be twice as high as the fraction of elements in the log spectra. By extension, the fraction of missing elements for double differenced log spectra can be four times as high as that for log spectra. These problems are not encountered using the methods described in this paper, since the log-spectral sequence is reconstructed prior to using it for differencing purposes.

4. SUMMARY AND CONCLUSIONS

In this paper we describe a series of new ways to recover missing feature information from log-spectral representations of speech. These compensation algorithms modify the incoming feature vector without any changes to the speech recognition system, in contrast to previously-described approaches based on missing-feature reconstruction (*e.g.* [2, 3]) or multi-channel recognition (*e.g.* [4, 5]). We describe two types of algorithms. The first approach, cluster-based inference, clusters the log-spectral vectors representing clean speech, and recovers missing data by estimating the spectral cluster in each analysis frame on the basis of the features that are present. The second approach, correlation-based inference, uses MAP procedures to estimate the values of missing data elements based on their correlation with the features that are present. Greatest recognition accuracy was obtained using the correlation-based inference approach, because, we believe, of its ability to exploit the temporal as well as spectral structure of speech. The recognition accuracy provided by our algorithms approaches but does not exceed that obtained by traditional mean imputation. Nevertheless, we believe that our algorithms provide greater computational efficiency and enable greater flexibility in recognition architecture.

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. Gales, M. and Young, S. (1995) “A fast and efficient implementation of Parallel Model Combination”, Proc. ICASSP 1995