ROBUST SPEECH RECOGNITION: THE CASE FOR RESTORING MISSING FEATURES

Bhiksha Raj[1], Michael L. Seltzer[2], and Richard M. Stern[2]
1. Mitsubishi Electric Research Labs
Cambridge, MA 02139 USA
2. Department of Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA

ABSTRACT

Speech recognition systems perform poorly in the presence of corrupting noise. Missing feature methods attempt to compensate for the noise by removing unreliable noise corrupted components of a spectrographic representation of the noisy speech and performing recognition with the remaining reliable components. Conventional classifier-compensation methods modify the recognition system to work with the incomplete representation so obtained. This constrains them to perform recognition using spectrographic features which are known to be suboptimal to cepstra. In previous work we have proposed an alternative feature-compensation approach whereby the unreliable components are replaced by estimates derived from the reliable components and the known statistics of clean speech. In this paper we perform a detailed comparison of various aspects of classifier-based and feature-based compensation methods. We show that although the classifier-based compensation methods are superior when recognition is performed with spectrographic features, feature-based compensation methods provide better recognition performance overall, since cepstra derived from the reconstructed spectrogram can now be used for recognition. In addition, they have the added advantages of being computationally less expensive and not requiring modification of the recognizer.

1. INTRODUCTION

Speech recognition systems perform poorly when the speech being recognized has been corrupted by noise. Missing-feature approaches comprise one family of noise compensation algorithms that have shown an ability to provide highly robust recognition in the presence of high levels of noise. In these approaches noise-corrupted regions of a spectrographic representation of the speech signal are identified and deemed unreliable, and recognition is performed with the remaining reliable regions of the spectrogram.

Most current techniques using missing features (*e.g.* [1][2]) are based on modifying the manner in which the recognition system computes likelihoods of classes or states to account for unreliable features. In class-conditional imputation the most likely values for the unreliable components of a vector for any class are used in computing the likelihood of that class. In marginalization, unreliable components are integrated out of the class densities prior to computing likelihoods. In more recent work [3], the hard binary decisions that determine the reliability or unreliability of components have been replaced by soft decisions which associate a score between 0 and 1. Characterizing each spectrographic element by its degree of reliability in this fashion results in improved recognition performance. We refer to these approaches as *classifier-compensation methods* since compensation for unreliable data is performed within the classifier (or recognizer).

In classifier-compensation methods the recognizer must explicitly model the distribution of the feature vectors with unreliable components and recognition must be performed using these features. Typically, these spectrographic features are log Mel spectra. It is well known, however, that log spectra are a suboptimal feature domain for recognition and that cepstral coefficients derived from log spectra typically provide significantly greater recognition accuracy. In fact, in some cases using noisy cepstral coefficients results in higher recognition accuracy than the use of log spectra derived from clean speech.

As an alternative approach, we have proposed in previous work the use of missing-feature methods that provide robust recognition through *feature compensation* [4,5]. These methods modify the incoming features rather than the manner in which recognition is performed. The unreliable log spectral components are erased and reconstructed using statistical information derived from clean speech and the remaining reliable components. This provides a complete set of log spectral vectors from which standard cepstral coefficients can be derived. This approach has two distinct advantages over classifier-compensation methods: compensation can be performed without modifying a standard speech recognition system, and recognition can be performed in the cepstral domain, resulting in greater recognition accuracy.

While both classifier-compensation and feature-compensation methods are effective, they differ in several aspects including robustness to errors in identifying noisy elements, the effects of additional processing of the log spectra, and computational complexity. In this paper we perform a detailed quantitative comparison of several aspects of classifier-compensation and feature-compensation methods. For simplicity we restrict ourselves at present to the use of binary decisions of reliability or unreliability. We also compare results to those of an alternative combination approach [2] wherein distributions of HMM states (in an HMM-based recognizer) hypothesized by classifier-compensation are used to reconstruct unreliable elements. On the basis of our comparisons we believe that feature-compensation methods are superior to the other methods in most aspects at noise levels that are typical in normal applications.

In Section 2 of this paper we briefly describe the various missing-feature methods. In Section 3 we describe our experimental setup. In Section 4 we discuss identification of unreliable components of spectrograms. In Section 5 we analyze the effect of data and feature preprocessing on the various methods. In Sec-

tion 6 we evaluate the overall recognition performance of the various methods. In Section 7 we compare the computational complexity of the various missing-feature methods. Finally in Section 8 we present our conclusions.

2. MISSING-FEATURE METHODS

In this section we very briefly review the various missing-feature methods.

2.1. Classifier-compensation methods

In classifier-compensation methods (*e.g.* [1,2]) the basic manner in which the likelihood of a class is computed is modified. There are two ways of doing this:

Class-conditional imputation: In this method, when computing the likelihood of any class or state (for an HMM-based recognizer), unreliable components of a log-spectral vector are replaced by their MAP estimates given the prior distribution of that class or state. These MAP estimates are then used to compute the likelihood of that class or state.

Marginalization: In this method the unreliable components of a log-spectral vector are integrated out of the distribution of a class, constrained by upper and lower bounds on the true values of these components implicit in their observed values. The resulting distributions with the smaller number of components are then used to compute the likelihood for that vector.

2.2. Feature-compensation methods

In feature-compensation methods (*e.g.* [5]) the unreliable components of the spectrogram are estimated based on the reliable components and the known statistical properties of log spectra. Recognition can then be performed either with the complete log spectra so derived, or with cepstra derived from them.

Cluster-based reconstruction: Here the log-spectral vectors of clean speech are first clustered. To estimate the unreliable components of any log spectral vector, the cluster that the vector belongs to is identified based only on its reliable components. The distribution of that cluster is then used to obtain MAP estimates for the unreliable components.

Covariance-based reconstruction: Here the probability distributions of component feature vectors in a spectrogram are assumed to be stationary, and correlations between any two such vectors are learned from spectrograms of clean speech. MAP estimates of unreliable elements of noisy spectrograms are obtained based on their correlations with reliable elements, assuming that the underlying distributions are Gaussian.

3. EXPERIMENTAL SETUP

The DARPA Resource Management (RM) database and the SPHINX-III HMM-based speech recognition system were used in all the experiments described in this paper. Context-dependent HMMs with 2000 tied states, each modelled by a single Gaussian, were trained using both the log spectra and cepstra of clean speech. The language weight was kept to a minimum in all cases, in order to emphasize the effect of the noisy acoustics on recognition accuracy. Mean normalization of features was performed in all experiments except those involving marginalization. Test utterances were corrupted with white noise and randomly-chosen samples of music from the Marketplace news program, as appropriate. In all cases both the additive noise and the clean speech samples were available separately, making it possible to evaluate the true SNR of any element in the spectrogram of the noisy utterances.

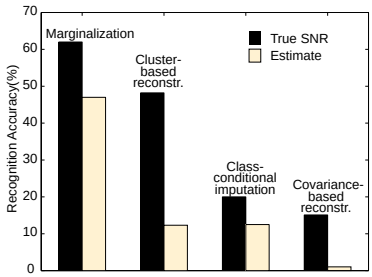

Figure 1. Comparisons of recognition accuracy obtained when unreliable elements are identified based on knowledge of their true SNR with accuracy obtained when the positions of unreliable elements must be estimated.

4. IDENTIFYING UNRELIABLE ELEMENTS OF THE SPECTROGRAM

An unreliable component of a spectrogram is generally defined as one with a local SNR that lies below a threshold. The optimal value of this threshold is dependent on the method used and was empirically found to be about –5 dB for feature-compensation methods and class-conditional imputation and about 15 dB for marginalization.

For missing-feature methods to be practicable, the unreliable components must be identified without *a priori* knowledge of the true SNR of spectrographic elements. Conventionally this is done by maintaining a running estimate of the noise spectrum and using this to estimate which elements of the spectrogram are unreliable. This method has the disadvantage of requiring that the spectrum of the corrupting noise be estimated, a problem that is almost intractable when the noise is non-stationary.

In this paper we chose to use a classifier-based method to identify noisy elements of the spectrogram. This reduces the task of identifying unreliable spectrogram elements from SNR estimation to a simpler binary decision process. The features used in classification were designed to exploit the characteristics of the speech signal itself. Two of the features, used for voiced speech segments characterize the harmonicity and periodicity often present in the signal. Four additional features, used for both voiced and unvoiced speech, capture information about the subband energy levels and spectral contour across frequency. Details of the mask-estimation classifier can be found in [6].

The effect of errors in identifying unreliable elements can be different for the different missing-feature methods. Figure 1 shows recognition accuracies obtained for several missing-feature methods applied to speech corrupted by white noise to 10 dB. We compare recognition accuracy obtained using perfect "oracle" knowledge of the true SNR values of spectrographic elements to identify unreliable feature locations with the corresponding accuracy obtained when the decisions about locations of unreliable elements must be obtained blindly from noisy data.

Marginalization shows the greatest robustness to errors in estimation of unreliable elements. In general, the classifier-compensation methods are much more robust to errors than the feature-

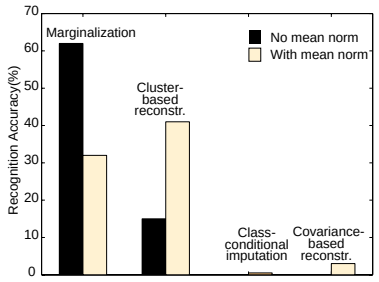

Figure 2. Recognition accuracy using various missing feature methods for speech in white noise at 10 dB SNR, with and without mean normalization of the features. The accuracy obtained with class-conditional imputation and covariance-based reconstruction without mean normalization is 0%.

compensation methods.

5. PREPROCESSING

Speech recognition systems do not normally use the features derived from incoming speech data directly for recognition. Some preprocessing of the incoming data is usually performed. The feature vectors for the utterance are usually "mean normalized", *i.e.* the mean value of the feature vectors is subtracted from all the vectors. This is known to result in a relative improvement in the word error rate by up to 25%.

When missing-feature methods are applied, however, it is not clear whether this procedure is useful. In Fig. 2. we show the effect of mean normalization on the recognition accuracy obtained with various missing-feature methods on speech corrupted to 10 dB by white noise. Unreliable spectrogram elements were identified using perfect knowledge of the true SNR in each case. Both reliable and unreliable components were used in computing the mean value of the vectors.

We observe that mean normalization is useful in all cases where estimation of unreliable components is performed. For marginalization, however, mean normalization actually results in a degradation of performance.

The incoming speech data itself may also be processed, *e.g.* by spectral subtraction [7] to reduce the noise level in the signal. This is likely to be helpful even for missing-feature methods since the components of the spectrogram that have been termed reliable still contain some amount of noise.

Figure 3 shows the recognition accuracy obtained with various missing-feature methods on speech corrupted to 10 dB by white noise. Recognition was performed with log spectra in all cases

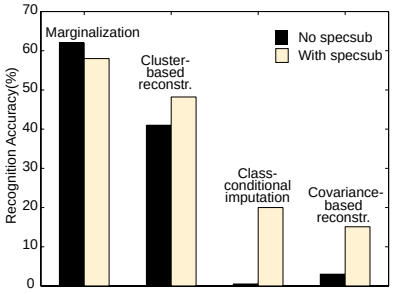

Figure 3. Recognition accuracy of various missing feature methods on speech in white noise at 10 dB SNR, with and without spectral subtraction.

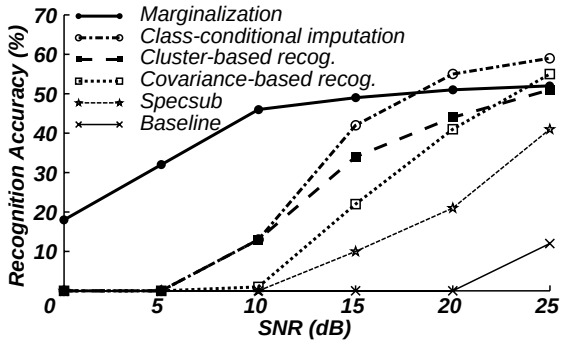

Figure 4. Recognition accuracy using various missing feature methods on speech corrupted by white noise. The baseline recognition accuracy obtained with no compensation is also shown.

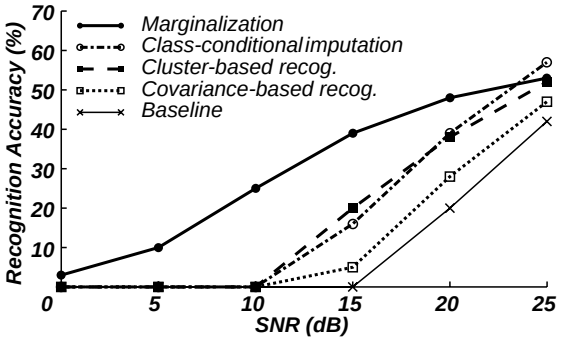

Figure 5. Recognition accuracy using various missing feature methods on speech corrupted by music. The baseline recognition accuracy obtained with no compensation is also shown.

and the true SNR of spectrographic elements was used to identify unreliable elements. We observe that spectral subtraction results in large improvements in recognition accuracy for feature-compensation methods. For marginalization, however, there is no noticeable improvement, presumably because the noise level in the reliable elements is very low to begin with. This is an advantage for marginalization as additional noise compensation steps can be avoided.

6. OVERALL RECOGNITION PERFORMANCE

Figures 4 and 5 show the recognition accuracies obtained with the various missing-feature methods on speech corrupted to several SNRs by white noise and music respectively. Recognition was performed using log spectra in all cases. Spectral subtraction and mean normalization were performed for all methods except marginalization. The location of unreliable elements was estimated. The best recognition accuracies are obtained using marginalization. This is to be expected since marginalization performs optimal classification.

As noted before, the advantage with feature compensation is that the reconstructed log spectra can be transformed to cepstra and used for recognition, something that is not possible with the classifier-compensation methods.

Figure 6 shows the corresponding recognition accuracy obtained with cepstra derived from the reconstructed log spectra. Compar-

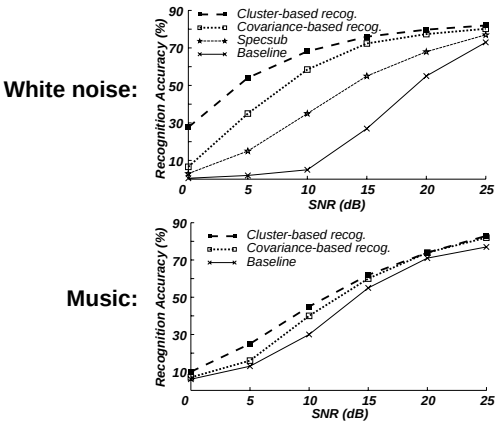

Figure 6. Recognition accuracy obtained with cepstra derived from log spectra reconstructed using feature-compensation methods for speech corrupted by white noise and music.

ison of Figures 4, 5 and 6 shows that the ability to perform cepstra-based recognition easily outweighs the advantages due to the optimal classification and those due to the greater robustness to errors in estimating unreliable elements that are characteristic of marginalization. The advantage however diminishes as the SNR decreases to 0 dB or so.

One could combine the classifier-compensation and feature-compensation methods by using the distributions of the HMM-state sequence hypothesized by a classifier-compensation method to reconstruct unreliable elements [2]. Figure 7 shows the recognition accuracy obtained with cepstra derived from log spectra reconstructed using both class-conditional imputation and marginalization in this manner. We note that overall, these methods are not more effective than feature-compensation methods

7. COMPUTATIONAL COMPLEXITY

The computational complexity of the various missing-feature methods also varies. Marginalization requires the computation of an error function for every unreliable component of a vector for every Gaussian in every HMM state considered. Cluster-based reconstruction similarly requires computation of error functions for every unreliable component for every cluster in the cluster-based representation. Class-conditional imputation and covariance-based reconstruction, on the other hand, only require MAP estimation of unreliable elements. Figure 8 shows the average time in seconds taken by a 400-MHz DEC alpha to recognize an utterance of speech corrupted to 10 dB by white noise, using the various missing-feature methods. This includes the time taken for computation of log spectra, reconstruction of unreliable elements, and transformation to cepstra in the case of the feature-compensation methods, and recognition. The time taken for identifying unreliable elements is not included. Marginalization is by far the most expensive of the methods. Feature-compensation methods do not generally increase the time taken for recognition significantly over the baseline.

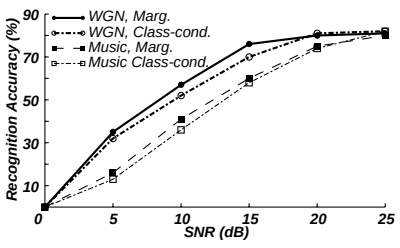

Figure 7. Recognition accuracy using cepstra derived from log spectra reconstructed using state sequences hypothesized by classifier-compensation methods. Results are shown for speech corrupted both my white noise and music.

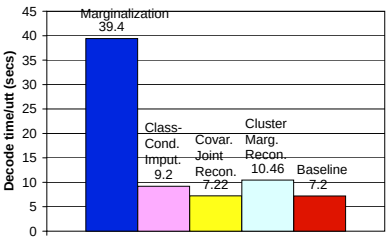

Figure 8. Average time in seconds needed to recognize an utterance using different missing-feature methods.

8. CONCLUSIONS

Missing-feature methods are generally very effective in compensating for both stationary and non-stationary noises. Of these, classifier-compensation methods such as marginalization are clearly superior when recognition is performed with spectrographic features. However, feature-compensation methods permit derivation of cepstral features, which result in better recognition accuracies than the best classifier-compensation method at most SNRs that are encountered in typical operating conditions. In addition, they are computationally less expensive and do not require any modification of the recognizer; the feature-compensation module can simply be used as a preprocessing block for any standard recognition system. The results of this paper have been obtained using binary decisions of whether spectrogram elements are reliable or unreliable. It is known that better results can be obtained by taking soft decisions [3]. However, since soft decisions can also be used in feature-compensation methods, we do not expect this to affect the conclusions stated above.

REFERENCES

1. 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.
2. L. Josifovski, M. Cooke, P. Green and A. Vizinho, “State based imputation of missing data for robust speech recognition and speech enhancement”, Proc. Eurospeech 1999.
3. J. Barker, L. Josifovski, M.P. Cooke and P.D. Green, “Soft decisions in missing data techniques for robust automatic speech recognition”, Proc. ICSLP 2000.
4. Raj, B., Singh R., and Stern, R.M. (1998) “Inference of missing spectrographic features for robust speech recognition”, Proc. ICSLP 1998.
5. Raj, B., Seltzer, M., and Stern, R. M., ``Reconstruction of damaged spectrographic features for robust speech recognition,'' Proc. ICSLP 2000.
6. Seltzer, M., Raj, B., and Stern, R. M., “Classifier-based mask estimation for missing feature methods of robust speech recognition”, Proc. ICSLP 2000.
7. Boll, S.F. (1979), “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Transactions on Acoustics, Speech and Signal Processing, April, 1979.