Available online at www.sciencedirect.com

SCIENCE DIRECT®

Speech Communication 43 (2004) 275–296

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Reconstruction of missing features for robust speech recognition

Bhiksha Raj [a,*], Michael L. Seltzer [b], Richard M. Stern [b]

[a] *Mitsubishi Electric Research Labs, 201 Broadway, 8th Floor, Cambridge, MA 02139, USA*
[b] *Carnegie Mellon University, Pittsburgh, PA 15213, USA*

Received 24 May 2002; received in revised form 9 May 2003; accepted 31 March 2004

Abstract

Speech recognition systems perform poorly in the presence of corrupting noise. Missing feature methods attempt to compensate for the noise by removing noise corrupted components of spectrographic representations of noisy speech and performing recognition with the remaining reliable components. Conventional classifier-compensation methods modify the recognition system to work with the incomplete representations so obtained. This constrains them to perform recognition using spectrographic features which are known to be less optimal than cepstra. In this paper we present two missing-feature algorithms that reconstruct complete spectrograms from incomplete noisy ones. Cepstral vectors can now be derived from the reconstructed spectrograms for recognition. The first algorithm uses MAP procedures to estimate corrupt components from their correlations with reliable components. The second algorithm clusters spectral vectors of clean speech. Corrupt components of noisy speech are estimated from the distribution of the cluster that the analysis frame is identified with. Experiments show that, although conventional classifier-compensation methods are superior when recognition is performed with spectrographic features, cepstra derived from the reconstructed spectrograms result in better recognition performance overall. The proposed methods are also less expensive computationally and do not require modification of the recognizer.

1. Introduction

Automatic speech recognition (ASR) systems perform poorly when the speech to be recognized is corrupted by noise, especially when the system has been trained on clean speech. Several algorithms have been proposed in the literature to compensate for the effects of noise on ASR systems. Most of these algorithms attempt to characterize the noise and model its effects on the speech signal explicitly (e.g. Varga and Moore, 1990; Acero, 1993; Gales and Young, 1996; Moreno, 1996) in order to compensate for it. The performance of these algorithms is usually critically

[*] Corresponding author. Tel.: +1 617 621 7593; fax: +1 617 621 7550.
E-mail address: bhiksha@merl.com (B. Raj).

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.03.007

dependent on the ability to measure the noise characteristics accurately, and they frequently fail to be effective when such measurement is difficult, such as when the noise is non-stationary (Raj et al., 1997).

In the mid-1990s researchers at the university of Sheffield proposed an alternative approach to noise compensation, the *missing feature* approach, that is based on exploitation of the inherent redundancy in the speech signal, rather than on explicit characterization of the noise (Cooke et al., 1994a). Speech signals have a large degree of redundancy built into them. For instance, speech that has been either high-pass filtered or low-pass filtered with a cutoff frequency of 1800 Hz remains perfectly intelligible (Fletcher, 1953). Similarly, speech that has undergone excision of spectral bands (Warren et al., 1995) or short temporal regions (Miller and Licklider, 1950) remains intelligible. Hence, one may hope to recognize speech effectively using only a fraction of the spectro-temporal information in the speech signal. To exploit this fact, missing feature methods represent speech using spectrographic time-frequency representations (that we will refer to as spectrograms in this paper), that consist of sequences of power spectral or log spectral vectors (which we generically refer to as spectral vectors in the rest of this paper). When the speech is corrupted by noise, some of the time-frequency components of this representation are more corrupted than others. The missing feature approach deems low-SNR time-frequency components as unreliable, and recognition is performed using only the remaining reliable components. The unreliable components are thus effectively assumed to be missing, and only the incomplete spectrographic information represented by the reliable components is assumed to be available.

In the two original algorithms proposed by the Sheffield group, recognition was performed by HMM-based recognizers directly with the incomplete spectrographic information from the reliable time-frequency components (Cooke et al., 1997). Since conventional HMM-based recognizers cannot perform recognition with incomplete representations, their algorithms modified the manner in which state output probabilities were computed within the recognizer. In the first algorithm, referred to as *state-based imputation*, computation of the output probability of any spectral vector, for any state, is accomplished by replacing unreliable components of the vector by maximum a posteriori (MAP) or minimum mean squared error (MMSE) estimates obtained from the reliable time-frequency components, computed from the distribution of that state. In the second algorithm, referred to as *marginalization*, the unreliable components are integrated out of the state output distributions. This latter approach is equivalent to the optimal classifier or recognizer, given the incomplete data. Later improvements to the algorithms incorporated the assumption that the value of any unreliable time-frequency component represents an upper bound on the *true* value of that component, i.e. the value that component would have had in the absence of corrupting noise, when the noise is additive and uncorrelated to the speech. This places an upper bound on the estimates of the unreliable components for state-based imputation (Josifovski et al., 1999). For marginalization, this places an upper limit on the integral that must be computed to marginalize out unreliable components from class distributions (Cooke et al., 2001). Since these methods modify the recognizer itself, we refer to them as *classifier-compensation* methods in this paper.

While both state-based imputation and marginalization have been shown to be extremely effective in compensating for noise, they suffer from several drawbacks. For them to be applicable, the state output distributions of the recognizer must represent the distributions of the spectral vectors where the reliable and unreliable components are identified. Recognition must therefore be performed with spectral vectors. However, speech recognition performance obtained using *cepstral* vectors has been found to be significantly superior to that obtained with spectral vectors (Davis and Mermelstein, 1980). It is infeasible to perform state-based imputation or marginalization effectively on cepstra-based recognizers since the distributions of spectral vectors cannot be derived from those of the lower-dimensional cepstral vectors. Another important drawback is that the recognizer must be modified to implement these algorithms. As a result, they can only be used in situations where one has access to the internals of the recognizer. There are other less serious problems as well. Utterance-level preprocessing steps, such as mean and variance normalization, that are known to improve recognition performance, cannot be performed with incomplete spectrographic data. The use of difference and double difference features, though possible, becomes more difficult and less effective. All these problems arise from the fact that these are classifier-compensation methods that attempt to perform recognition directly with the incomplete spectrograms, modifying the recognizer to account for the missing components.

In this paper we present two missing-feature algorithms that take an alternative approach: they *reconstruct complete spectrograms from the incomplete ones prior to recognition*. To achieve this, the true values of the unreliable time-frequency components of the spectrogram are estimated from the reliable components and the known statistical relationships between the various components of the spectrogram. Cepstral vectors can now be derived from the spectral vectors in the reconstructed spectrograms, for recognition. Utterance level processing such as mean normalization can also be performed. Recognition performance with the normalized cepstral vectors so obtained is frequently much better than that obtained by marginalization, which performs optimal recognition based on incomplete spectral vectors. Equally importantly, the recognizer itself need not be modified in any manner. This permits the usage of any form of recognizer, including off-the-shelf commercial recognizers that can take cepstral vectors as input. Since these algorithms work only on incoming feature vectors, we refer to them as *feature-compensation* methods.

Feature-compensation missing-feature algorithms have previously been reported by other researchers. Most algorithms model the distribution of the spectral vectors of clean speech as Gaussian mixtures. Dupont (1998) and Raj et al. (1998) compute a posteriori probabilities of all the Gaussians in the mixture from the reliable components of spectral vectors, ignoring unreliable components altogether. These probabilities are then used for MMSE estimation of unreliable components. Renevey (2001) adapts the parameters of the Gaussians in the mixture to the noise conditions of the speech to be recognized, using explicit characterizations of the noise distributions. A posteriori probabilities of all Gaussians in the mixture are computed using the modified distributions and used to obtain MMSE estimates of unreliable components.

In contrast, the two algorithms reported in this paper do not require explicit characterization of distributions of the noise. Further, they utilize information from unreliable spectrographic components by assuming that their observed values are upper bounds on their true values. *Correlation-based reconstruction* is based on a simple statistical model that represents the sequence of spectral vectors in the spectrogram as the output of a stationary Gaussian random process. A bounded version of the MAP estimation procedure is used to estimate unreliable components, based on the statistical parameters of this process. *Cluster-based reconstruction* is based on the more conventional Gaussian mixture representations of the distributions of clean speech. The reconstruction uses the bounded MAP estimation procedure to obtain Gaussian specific estimates of unreliable components, which are then combined into a final estimate.

A crucial component of missing feature methods is the identification of unreliable components in the spectrograms. Several solutions have been proposed for this problem in the literature (e.g. Cooke et al., 1994a,b; Drygajlo and El-Maliki, 1998; Vizhinho et al., 1999; Renevey and Drygajlo, 1999). These methods can be largely categorized into two: those that are derived from computational auditory scene analysis of the signal, and those that depend in some manner on tracking or measuring the corrupting noise. In this paper we treat the identification of noisy components as a Bayesian classification problem. This algorithm does not depend on characterizations of the distributions of the noise, performing classification based only on features measured from the noisy signal instead. We only provide a brief outline of the algorithm used in this paper. Additional details of the algorithm are presented in a companion paper (Seltzer et al., 2004).

The rest of this paper is arranged as follows: in Section 2 we give a brief description of the spectrographic representation used in the missing feature work described in this paper. In Section 3 we define some notations used in the rest of the paper. In Section 4 we briefly describe conventional state-based imputation and marginalization. In Section 5 we describe the proposed algorithms: covariance-based reconstruction and cluster-based reconstruction. In Section 6 we outline the method used to identify noise-corrupted components of the spectrogram. In Section 7 we describe several experimental results. Finally, in Section 8 we present our conclusions.

2. Spectrographic representations

In all of our work the spectrographic representation used for the speech signal has been the Mel spectral representation (O'Shaughnessy, 1987), or the Mel spectrogram. This consists of a sequence of *Mel log-spectral* vectors, each of which represents the frequency warped log spectrum of a short frame of speech, typically 20 ms wide. Fig. 1a shows the Mel spectrogram of a typical clean speech signal.

Additive noise affects different regions of the Mel spectrogram differently. Fig. 1b shows the Mel spectrogram of the signal in Fig. 1a, when it has been corrupted to 10 dB by white noise. Comparison of the two figures shows that while some regions are relatively unaffected by the noise, others are badly corrupted. The degree of corruption of any time-frequency component of the spectrogram is dependent on the SNR of that component. Missing feature methods assume that the effect of the noise is to render all low-SNR regions unreliable. Thus, all time-frequency components that have an SNR below a particular threshold are assumed to be unreliable. However, the values of these unreliable components are assumed to be the upper bound on their *true* values, i.e. the value that they would have had in the absence of corrupting noise. This is based on the assumption that the noise is additive and uncorrelated to the speech. All time-frequency components whose SNR lies above the threshold are assumed to be reliable, and good approximations to their true values. The optimal value of the threshold is different for different missing-feature methods, and also varies with the global SNR of the noisy signal. In general, however, the threshold −5 dB was empirically found to be close to optimal across a wide variety of SNRs for the methods reported in this paper and for state-based imputation. For marginalization, the optimal threshold was found to be 0 dB.

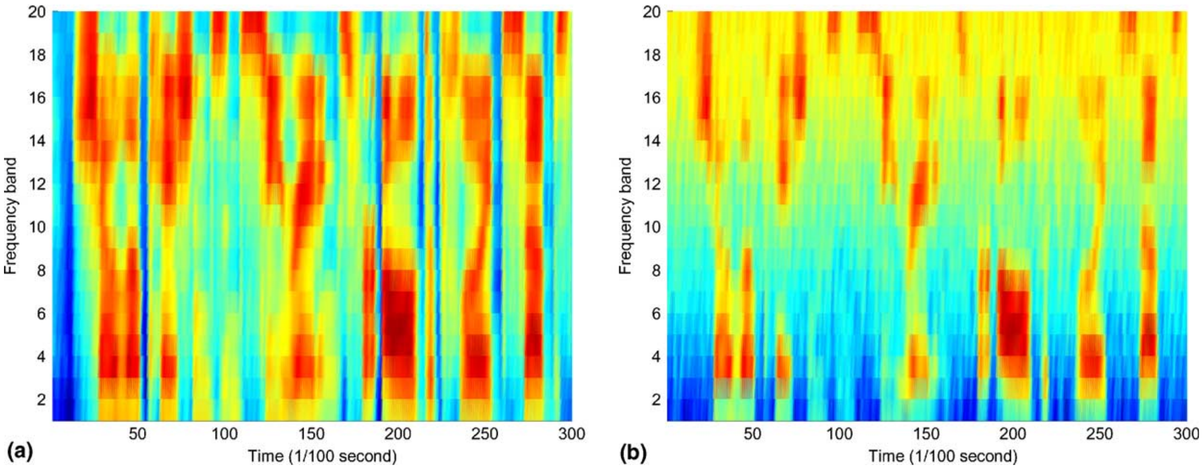


Fig. 1. (a) Mel spectrogram of a clean speech signal. The utterance is "show locations and C-ratings of all deployed subs". (b) Mel spectrogram of the same signal when it has been corrupted to 10 dB by white noise.

3. Notation

Before proceeding, we establish some of the notation and terminology used in the rest of the paper. Every frame of incoming speech has underlying clean speech that has been corrupted by noise to result in the observed noisy speech. Corresponding to the tth frame of noisy speech, there is a measured noisy spectral vector $Y(t)$. The vector $Y(t)$ has a set of reliable components, that we arrange into a vector $Y_r(t)$ and a set of unreliable components, which we arrange into the vector $Y_u(t)$. We refer to $Y_u(t)$ as the *unreliable component vector* of $Y(t)$ and to $Y_r(t)$ as the *reliable component vector* of $Y(t)$. $Y(t)$ is a union of the two vectors. We can express the relation between $Y(t)$, $Y_r(t)$ and $Y_u(t)$ as

$$\begin{aligned} Y_{\rm r}(t) &= R(t)Y(t) \\ Y_{\rm u}(t) &= U(t)Y(t) \\ Y(t) &= A(t)[Y_{\rm r}(t)^{\rm T}Y_{\rm u}(t)^{\rm T}]^{\rm T} \end{aligned} \tag{1}$$

where $R(t)$ and $U(t)$ are permutation matrices that select the reliable and unreliable components respectively of $Y(t)$ and arrange them into $Y_r(t)$ and $Y_u(t)$, the superscripted 'T' represents transposition, $[Y_r(t)^T Y_u(t)^T]^T$ is a vector constructed by concatenating the transposes of $Y_r(t)$ and $Y_u(t)$, and $A(t)$ is the permutation matrix that rearranges the components of $[Y_r(t)^T Y_u(t)^T]^T$ to give $Y(t)$.

Corresponding to the noisy spectrogram, i.e. the spectrogram of the noisy speech, is a *true* spectrogram which is the spectrogram that would have been computed had the signal not been corrupted by noise. Corresponding to each noisy spectral vector $Y(t)$ from the noisy spectrogram, there is a *true* spectral vector $X(t)$ from the true spectrogram. The components of $X(t)$ that correspond to the reliable and unreliable components of $Y(t)$ can also be arranged into vectors $X_r(t)$ and $X_u(t)$. $X_r(t)$ and $X_u(t)$ are related to $Y_r(t)$ and $Y_u(t)$ as follows:

$$\begin{aligned} X_{\rm r}(t) &\approx Y_{\rm r}(t) \\ X_{\rm u}(t) &\leqslant Y_{\rm u}(t) \end{aligned} \tag{2}$$

We refer to the components of $X_u(t)$ as the *unreliably known* components of $X(t)$, since their value is not known, and to X_u(t) as the *unreliably known component vector* of $X(t)$. Similarly we refer to the components of $X_r(t)$ as the *reliably known* components of $X(t)$, and to $X_r(t)$ as the *reliably known component vector* of $X(t)$.

4. Classifier-compensation methods

In this section we briefly describe how state-based imputation and marginalization modify the computation of state output probabilities in HMM-based speech recognition systems. Both algorithms have been well documented in various papers, and we only recapitulate the salient points here for reference. For more detailed information the reader is referred to the several papers on the subject (e.g. Lippmann and Carlson, 1997; Cooke et al., 2001).

4.1. State-based imputation

In most HMM-based systems state output probabilities are modelled as mixtures of Gaussians. For any vector $X(t)$ with reliably known component vector $X_r(t)$ and unreliably known component vector $X_u(t)$, the state output probability of a state s, $P(X(t)|s)$, can be expressed as

$$\begin{aligned} P(X(t)|s) &= P(X_{\rm r}(t), X_{\rm u}(t)|s) \\ &= \sum_j c_{j,s} G(X_{\rm r}(t), X_{\rm u}(t); \mu_{j,s}, \Theta_{j,s}) \end{aligned} \tag{3}$$

where $G(X_r(t), X_u(t);\mu_{j,s}, \Theta_{j,s})$ represents the jth Gaussian in the mixture Gaussian density for s with mean vector $\mu_{j,s}$ and covariance matrix $\Theta_{j,s}$, and $c_{j,s}$ is the mixture weight of the jth Gaussian. For any noisy spectral vector $Y(t)$, one would ideally compute the state output probability of the underlying true vector $X(t)$. State-based imputation approximates this as

$$\begin{aligned} P(X(t)|s) &= P(Y_{\rm r}(t), \widehat{X}^s_{\rm u}(t)|s) \\ &= \sum_j c_{j,s} G(Y_{\rm r}(t), \widehat{X}^s_{\rm u}(t); \mu_{j,s}, \Theta_{j,s}) \end{aligned} \tag{4}$$

where $\widehat{X}^s_u(t)$ is an MMSE estimate of $X_u(t)$ obtained from $Y_r(t)$ and the output distribution of s (Josifovski et al., 1999), computed as

$$\widehat{X}^s_{\mathrm{u}}(t) = \sum_j \gamma_{j,s}(Y(t))U(t)\mu_{j,s} \quad (5)$$

where $U(t)$ is the permutation matrix that selects unreliable components from $Y(t)$ to form $Y_{\mathrm{u}}(t)$ and

$$\gamma_{j,s}(Y(t)) = \frac{c_{j,s}\int_{-\infty}^{Y_{\mathrm{u}}(t)} G(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t); \mu_{j,s}\Theta_{j,s})\,\mathrm{d}X_{\mathrm{u}}(t)}{\sum_k c_{k,s}\int_{-\infty}^{Y_{\mathrm{u}}(t)} G(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t); \mu_{k,s}, \Theta_{k,s})\,\mathrm{d}X_{\mathrm{u}}(t)} \quad (6)$$

While other forms of the estimate for $\widehat{X}^s_{\mathrm{u}}(t)$ have also been proposed (e.g. Renevey, 2001) the basic principle behind the implementation of the algorithm remains unchanged.

4.2. Marginalization

In marginalization, the unreliable components of the state distributions are simply integrated out of the state output distributions. State output probabilities are computed as $P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|s)$. When state output densities are modelled by mixtures of Gaussians the state output density value for state is computed as

$$P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|s) = \sum_k c_{k,s}\int_{-\infty}^{Y_{\mathrm{u}}(t)} G(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t); \mu_{k,s}, \Theta_{k,s})\,\mathrm{d}X_{\mathrm{u}}(t) \quad (7)$$

5. Feature-compensation methods

The methods described in Section 4 modify the recognizer in order to perform recognition with incomplete spectrographic information. In this section we present two new feature-compensation algorithms, *correlation-based reconstruction* and *cluster-based reconstruction*, that reconstruct complete spectrograms from the incomplete ones. These algorithms estimate the true value of the unreliable spectrographic components from the reliable components. The simplest method of estimating these values is by simple interpolation between the closest reliable components. However, as reported by Raj (2000), simple interpolation-based reconstruction is ineffective for spectrograms of noisy speech signals. Instead, the algorithms presented in this section estimate unreliable spectrographic components based on the known statistical properties of spectral vectors. We describe the algorithms in greater detail in the following subsections.

5.1. Correlation-based reconstruction

In correlation-based reconstruction the sequence of spectral vectors that constitute the spectrogram of a clean speech signal are considered to be the output of a Gaussian wide-sense stationary (WSS) random process (Papoulis, 1991). All clean speech spectrograms are assumed to be individual observations of the same process. The assumption of wide-sense stationarity implies that the means of the spectral vectors and the covariances between components of the spectrogram are independent of their position in the spectrogram. If we represent the mean of the kth component of the tth spectral vector $X(t, k)$ of an utterance as $\mu(t, k)$, and the covariance between the k_1th component of the t_1th spectral vector $X(t_1, k_1)$ and the k_2th component of the t_2th spectral vector $X(t_2, k_2)$ as $c(t_1, t_2, k_1, k_2)$, we have

$$\begin{aligned}\mu(t,k) &= E[X(t,k)] \\ c(t_1,t_2,k_1,k_2) &= E[(X(t_1,k_1) - \mu(t_1,k_1))(X(t_2,k_2) - \mu(t_2,k_2))]\end{aligned} \quad (8)$$

where $E[\,]$ stands for the expectation operator. The assumption of wide-sense stationarity gives us the following properties for these parameters:

$$\mu(t,k) = \mu(t_1,k) = \mu(k) \quad (9)$$

$$\begin{aligned}c(t,t+\tau,k_1,k_2) &= c(t_1,t_1+\tau,k_1,k_2) \\ &= c(\tau,k_1,k_2)\end{aligned} \quad (10)$$

In other words, the expected value $\mu(k)$ of the kth component of a spectral vector is not dependent on where the vector occurs in the spectrogram. Similarly, the covariance between the components of two spectral vectors depends only on the distance τ between the vectors (along the time axis) and not on where the vectors occur in the spectrogram. The relative covariance $r(t_1, t_1 + \tau, k_1, k_2)$ between any two components $X(t_1, k_1)$ and $X(t_1 + \tau, k_2)$ is also dependent only on τ and is given by

$$r(t_1, t_1 + \tau, k_1, k_2) = r(\tau, k_1, k_2) = \frac{c(\tau, k_1, k_2)}{\sqrt{c(0, k_1, k_1)c(0, k_2, k_2)}} \tag{11}$$

The means of the components of the spectral vectors $\mu(k)$ and the various covariance parameters $c(\tau, k_1, k_2)$ can be learnt from the spectrograms of a training corpus of clean speech. Let $X^j(t, k)$ represent the kth component of the tth spectral vector from the jth training signal. The various mean and covariance values can be estimated as

$$\mu(k) = \frac{1}{\sum_j N_j} \sum_j \sum_t X^j(t,k)$$
$$c(\tau, k_1, k_2) = \frac{1}{\sum_j (N_j - \tau)} \sum_j \sum_t (X^j(t, k_1) - \mu(k_1)) \times (X^j(t + \tau, k_2) - \mu(k_2)) \tag{12}$$

Relative covariance values can be computed from the covariance values using Eq. (11). The implication of the assumption of a Gaussian process is that the joint distribution the components of all the spectral vectors in a sequence of vectors is assumed Gaussian. Consequently, the distribution of any subset of these components is also Gaussian (Papoulis, 1991). Thus, the estimated mean and covariance values characterize the process completely and no other statistical parameters need be estimated.

The task of reconstruction is to reconstruct the underlying true spectral vector for every spectral vector in the noisy spectrogram. Let $Y(t)$ be the noisy spectral vector whose true counterpart $X(t)$ must be reconstructed. As before, let $Y_u(t)$ and $Y_r(t)$ be the unreliable and reliable component vectors of $Y(t)$, $X_u(t)$ and $X_r(t)$ the corresponding counterparts from $X(t)$. $X_r(t)$ can be approximated by $Y_r(t)$. Only $X_u(t)$ must be estimated to reconstruct $X(t)$ completely. We now construct a *neighborhood vector* $Y_n(t)$ from all reliable components of the spectrogram that have a relative covariance greater than a threshold value with at least one of the components of $X_u(t)$. Let $X_n(t)$ be the underlying true value of $Y_n(t)$. Since all the components of $Y_n(t)$ are reliable, $X_n(t) \approx Y_n(t)$. The joint distribution of $X_u(t)$ and $X_n(t)$ is Gaussian. The parameters of this distribution are the expected value of $X_u(t)$, $\mu_u(t)$, the expected value of $X_n(t)$, $\mu_n(t)$, the autocorrelation of $X_u(t)$, $C_{uu}(t)$, the autocorrelation of $X_n(t)$, $C_{nn}(t)$, and the cross correlation between $X_u(t)$ and $X_n(t)$, $C_{un}(t)$. These parameters can all be constructed from the mean and covariance terms learnt from the training corpus. Fig. 2 demonstrates the construction of $Y_u(t)$ and $Y_n(t)$ and the parameters of their joint distribution with an example. $X_u(t)$ is now estimated as

$$\widehat{X}_u(t) = \arg\max_{X_u}\{P(X_u(t)|X_n(t) = Y_n(t), X_u(t) \leqslant Y_u(t))\} \tag{13}$$

Denoting $X_n(t) = Y_n(t)$ as $Y_n(t)$ for simplicity, and using Bayes rule, this can be rewritten as

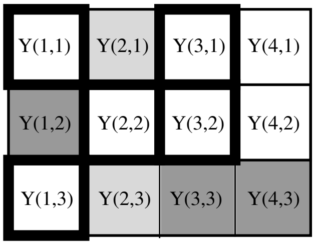
$$\widehat{X}_u(t) = \arg\max_{X_u}\{P(X_u(t), X_u(t) \leqslant Y_u(t)|Y_n(t))\} \tag{14}$$

We refer to the estimate given by Eq. (14) as a *bounded* MAP estimate. It can be shown that $P(X_u(t)|Y_n(t))$, the distribution of $X_u(t)$ conditioned on $X_n(t)$ being equal to $Y_n(t)$, is a Gaussian with mean $\mu(t) + C_{un}(t)C_{nn}^{-1}(t)(Y_n(t) - \mu_n(t))$. As shown in Appendix A, the solution to Eq. (14) can be obtained by the following iterative procedure:

Let $X_u(t, k)$ and $Y_u(t, k)$ be the kth components of $X_u(t)$ and $Y_u(t)$ respectively. Let the current estimate of $X_u(t, k)$ be $\overline{X}_u(t, k)$. The estimation procedure can now be stated as follows:

1. Initialize $\overline{X}_u(t,k) = Y_u(t,k)$, $1 \leqslant k \leqslant K$, where K is the total number of components in $X_u(t)$.
2. For each of the K components
   2a. Compute the MAP estimate

$$\widetilde{X}_u(t,k) = \arg\max_{X_u(t,k)}\{P(X_u(t,k)|Y_n(t), \overline{X}_u(t,j)\ \forall j,\ j \neq k)\} \tag{15}$$



Y(1,1)	Y(2,1)	Y(3,1)	Y(4,1)
Y(1,2)	Y(2,2)	Y(3,2)	Y(4,2)
Y(1,3)	Y(2,3)	Y(3,3)	Y(4,3)

Fig. 2. The figure represents a small spectrogram with four spectral vectors, each with four components. The grey components are missing. We wish to estimate all the missing components in the second spectral vector jointly. These are shown in a lighter shade of grey in the figure.[1]

This is simply the mean of the Gaussian distribution of $X_u(t,k)$, conditioned on the reliable values $Y_n(t)$ and on all other components of $X_u(t)$ being equal to their current estimates.

2b. Compute the bounded MAP estimate from the MAP estimate as

$$\overline{X}_u(t,k) = \min(\widetilde{X}_u(t,k), Y_u(t,k)) \tag{16}$$

3. If all $\overline{X}_u(t,k)$ estimates have converged, set $\widehat{X}_u(t,k) = \overline{X}_u(t,k)\ \forall k$ to obtain $X_u(t)$, else go back to Step 2.

$X_u(t)$ is estimated as described above for each spectral vector in the spectrogram. This, combined with the reliable components, reconstructs the entire spectrogram.

5.2. Cluster-based reconstruction

In cluster-based reconstruction the sequence of spectral vectors in the spectrogram is modelled as the output of an independent, identically distributed (IID) random process. Unreliable components of spectral vectors are reconstructed based on their statistical relationships to the reliable components from the same vector. This is in contrast to the assumptions behind correlation-based reconstruction, where all components of the spectrogram were assumed to be correlated and unreliable components of a vector were reconstructed based on their statistical relationship to reliable components in neighboring vectors as well.

In cluster-based reconstruction, the spectral vectors of clean speech are assumed to be segregated into a number of clusters. Each cluster is assumed to have a Gaussian distribution. The distribution of the kth cluster is thus given by

[1] $Y_u(2)$ is constructed as

$$Y_u(2) = [Y(2,1), Y(2,3)]^T$$

The neighborhood vector $Y_n(2)$ is constructed of all the components $Y(t,k)$, such that either $r(t-2,1,k) \geqslant 0.5$, or $r(t-2,3,k) \geqslant 0.5$. These are represented by the components with the thick outlines. This gives us

$$Y_n(2) = [Y(1,1), Y(1,3), Y(2,2), Y(3,1), Y(3,2)]^T$$

The mean vectors for $X_n(2)$ and $X_u(2)$, the clean speech counterparts of $Y_n(2)$ and $Y_u(2)$, are constructed as

$$E[X_n(2)] = \mu_n(2) = [\mu(1), \mu(3), \mu(2), \mu(1), \mu(2)]^T$$

$$E[X_u(2)] = \mu_u(2) = [\mu(1), \mu(3)]$$

The autocovariance matrix of $X_n(2)$ is a 5 × 5 matrix constructed as

$$C_{nn}(2) = \begin{bmatrix} c(0,1,1) & c(0,1,3) & c(1,1,2) & c(2,1,1) & c(2,1,2) \\ c(0,3,1) & c(0,3,3) & c(1,3,2) & c(2,3,1) & c(2,3,2) \\ c(-1,2,1) & c(-1,2,3) & c(0,2,2) & c(1,2,1) & c(1,2,2) \\ c(-2,1,1) & c(-2,1,3) & c(-1,1,2) & c(0,1,1) & c(0,1,2) \\ c(-2,2,1) & c(-2,2,3) & c(-1,2,2) & c(0,2,1) & c(0,2,2) \end{bmatrix}$$

The cross covariance between $X_u(2)$ and $X_n(2)$ is a 2 × 5 matrix constructed as

$$C_{un}(2) = \begin{bmatrix} c(-1,1,1) & c(-1,1,3) & c(0,1,2) & c(1,1,1) & c(1,1,2) \\ c(-1,3,1) & c(-1,3,3) & c(0,3,2) & c(1,3,1) & c(1,3,2) \end{bmatrix}^T$$

$$P(X|k) = \frac{\exp(-\frac{1}{2}(X-\boldsymbol{\mu}_k)^{\mathrm{T}}\Theta_k^{-1}(X-\boldsymbol{\mu}_k))}{\sqrt{(2\pi)^d|\Theta_k|}} \tag{17}$$

where X represents an arbitrary vector from the kth cluster, d represents the dimensionality of X, and μ_k and Θ_k represent the mean vector and covariance matrix of the kth cluster, respectively. The overall distribution of spectral vectors is thus a mixture Gaussian given by

$$\begin{aligned} P(X) &= \sum_{k=1}^{K} c_k P(X|k) \\ &= \sum_{k=1}^{K} \frac{c_k}{\sqrt{(2\pi)^d|\Theta_k|}} \\ &\quad \times \exp\left(-\frac{1}{2}(X-\boldsymbol{\mu}_k)^{\mathrm{T}}\Theta_k^{-1}(X-\boldsymbol{\mu}_k)\right) \end{aligned} \tag{18}$$

where c_k is the a priori probability of the kth cluster. The a priori probabilities, means, and covariances of the clusters must all be learnt from a training corpus. This can be done by explicitly clustering the spectral vectors of the training data using techniques such as the LBG algorithm (Linde et al., 1980) or k-means clustering (McQueen, 1967), and estimating the a priori probabilities and the distribution parameters of the individual clusters thereafter. In this paper, however, we compute all parameters jointly from the training corpus using the expectation maximization (EM) algorithm (Dempster et al., 1977).

The parameters learnt from the training data can be used to reconstruct the underlying true spectral vector for any noisy spectral vector. The unreliably known components of the true spectral vector can be estimated by determining the cluster to which the vector belongs, and estimating them from the distribution of that cluster. This concept is illustrated by Fig. 3. Identifying the correct cluster for any spectral vector is a classification problem. As always, this can be errorful, especially since the vectors are noisy and incomplete. To account for this, we obtain a separate estimate for the unreliable components from the distribution of *each* of clusters. This results in as many estimates as there are clusters. The final estimate is a weighted average of all the estimates, where the weight of any estimate obtained from the distribution of any cluster is the a posteriori probability of that cluster, given the reliable components of that vector.

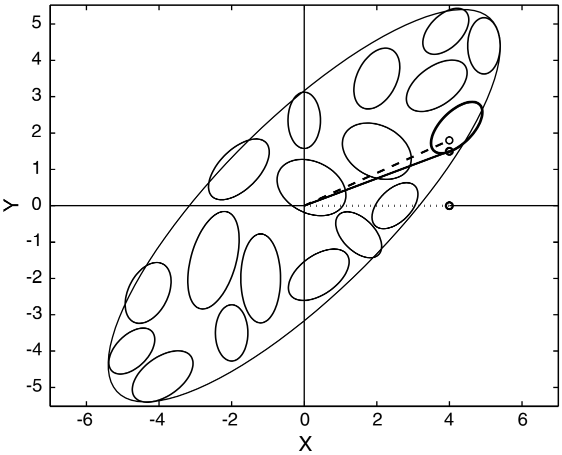


Fig. 3. Schematic representation of cluster-based reconstruction. The large ellipse represents the outline of the distribution of a set of two-dimensional vectors. The data has been segregated into a number of small clusters, represented by the smaller ellipses. The solid line represents a complete vector. The Y component of this vector is unreliable and only the X component, represented by the dotted line along the X axis, is reliably known. The cluster-based reconstruction method identifies the thick ellipse as the cluster that the complete vector belongs to, and uses the distribution of that cluster to obtain a bounded MAP estimate for the Y component, and thereby the complete vector, represented by the dashed line.

Let $Y(t)$ represent the noisy vector for which the underlying true vector $X(t)$ must be reconstructed. The reliably known component vector of $X(t)$, $X_{\mathrm{r}}(t)$, can be approximated by the reliable component vector of $Y(t)$, $Y_{\mathrm{r}}(t)$. The unreliably known component vector $X_{\mathrm{u}}(t)$ must be estimated. The estimate for $X_{\mathrm{u}}(t)$ obtained from the distribution of the kth cluster, $\widehat{X}^k_{\mathrm{u}}(t)$ is given by

$$\begin{aligned} \widehat{X}^k_{\mathrm{u}}(t) = \arg\max\nolimits_{X_{\mathrm{u}}}\{&P(X_{\mathrm{u}}(t)|k, X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t), \\ &X_{\mathrm{r}}(t) = Y_{\mathrm{r}}(t))\} \end{aligned} \tag{19}$$

where $P(X_{\mathrm{u}}(t)|k, X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t), X_{\mathrm{r}}(t) = Y_{\mathrm{r}}(t))$ is the distribution of $X_{\mathrm{u}}(t)$, conditioned on $X(t)$ belonging to the kth cluster, $X_{\mathrm{u}}(t)$ being no greater than $Y_{\mathrm{u}}(t)$, and $X_{\mathrm{r}}(t)$ being equal to $Y_{\mathrm{r}}(t)$. Using Bayes' rule and representing $X_{\mathrm{r}}(t) = Y_{\mathrm{r}}(t)$ simply as $Y_{\mathrm{r}}(t)$, this can be written as

$$\widehat{X}_{\mathrm{u}}^{k}(t) = \arg\max_{X_{\mathrm{u}}}\{P(X_{\mathrm{u}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|k, Y_{\mathrm{r}}(t))\} \tag{20}$$

The operation in Eq. (20) represents the bounded MAP estimation procedure described in Section 5.1. Since all cluster distributions are Gaussian, $P(X_{\mathrm{u}}(t)|k, Y_{\mathrm{r}}(t))$ is also Gaussian. The mean of the kth cluster, μ_k, can be partitioned into the two vectors $\mu_{k,u}(t)$, the expected value of $X_{\mathrm{u}}(t)$, and $\mu_{k,r}(t)$, which represents the means of the components of $X_{\mathrm{r}}(t)$. The components of the covariance matrix of the kth cluster, Θ_k, that correspond to $X_{\mathrm{u}}(t)$ and $X_{\mathrm{r}}(t)$ can be separated into $\Theta_{k,\mathrm{uu}}(t)$ and $\Theta_{k,\mathrm{rr}}(t)$ respectively. The cross-correlation between $X_{\mathrm{u}}(t)$ and, $X_{\mathrm{r}}(t)$, $\Theta_{k,\mathrm{ur}}(t)$ can also be derived from Θ_k. From these terms, the bounded MAP estimate of $X_{\mathrm{u}}(t)$ for the kth cluster can be obtained using the procedure described in Section 5.1 and Appendix A. The overall estimate of $X_{\mathrm{u}}(t)$ is given by

$$\widehat{X}_{\mathrm{u}}(t) = \sum_{k=1}^{K} P(k|Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t))\widehat{X}_{\mathrm{u}}^{k}(t) \tag{21}$$

where $P(k|Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t))$ is the a posteriori probability of the kth cluster and is given by

$$\begin{aligned} &P(k|Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)) \\ &\quad = \frac{c_k P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|k)}{\sum_{j=1}^{K} c_j P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|j)} \end{aligned} \tag{22}$$

In order to compute the term $P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|k)$, $P(X(t)|k)$, must be stated explicitly in terms of the reliably known and unreliably known component vectors of $X(t)$. This gives us

$$\begin{aligned} P(X(t)|k) &= P(X_{\mathrm{r}}(t), X_{\mathrm{u}}(t)|k) \\ &\quad P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|k) \\ &\quad = \int_{-\infty}^{Y_{\mathrm{u}}(t)} P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t)|k)\,\mathrm{d}X_{\mathrm{u}}(t) \end{aligned} \tag{23}$$

The term to the right is difficult to compute when $P(X(t)|k)$ is a Gaussian with non-zero off-diagonal elements in its covariance matrix. We therefore consider only the diagonal components of the covariance matrices when computing the a posteriori probabilities of clusters, assuming all other components to be 0. Under this assumption, the Gaussian distribution of the kth cluster can be expressed as

$$P(X(t)|k) = \prod_i \frac{\exp\left(-\frac{(X(t,i)-\boldsymbol{\mu}_k(i))^2}{2\theta_k(i)}\right)}{\sqrt{2\pi\theta_k(i)}} \tag{24}$$

where $\boldsymbol{\mu}_k(i)$ is the ith component of μ_k and $\theta_k(i)$ is the ith diagonal element of Θ_k. $P(X(t)|k)$ can now be separated out in terms of the reliably known and unreliably known components of $X(t)$ as

$$\begin{aligned} P(X(t)|k) = &\prod_{i|X(t,i)\in X_{\mathrm{r}}(t)} \frac{\exp\left(-\frac{(X(t,i)-\boldsymbol{\mu}_k(i))^2}{2\theta_k(i)}\right)}{\sqrt{2\pi\theta_k(i)}} \\ &\times \prod_{i|X(t,i)\in X_{\mathrm{u}}(t)} \frac{\exp\left(-\frac{(X(t,i)-\boldsymbol{\mu}_k(i))^2}{2\theta_k(i)}\right)}{\sqrt{2\pi\theta_k(i)}} \end{aligned} \tag{25}$$

The first product term in Eq. (25) computes the probabilities of all reliably known components of $X(t)$, i.e. all components of $X_{\mathrm{r}}(t)$, and the second product term computes the probability of all unreliably known components. $P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|k)$ can now be computed as

$$\begin{aligned} &P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|k) \\ &\quad = \prod_{i|X(t,i)\in X_{\mathrm{r}}(t)} \frac{\exp\left(-\frac{(Y(t,i)-\boldsymbol{\mu}_k(i))^2}{2\theta_k(i)}\right)}{\sqrt{2\pi\theta_k(i)}} \\ &\quad = \prod_{i|X(t,i)\in X_{\mathrm{u}}(t)} \int_{-\infty}^{Y(t,i)} \frac{\exp\left(-\frac{(X(t,i)-\boldsymbol{\mu}_k(i))^2}{2\theta_k(i)}\right)}{\sqrt{2\pi\theta_k(i)}}\,\mathrm{d}X(t,i) \end{aligned} \tag{26}$$

The a posteriori probabilities of the clusters, i.e. the $P(k|Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t))$ terms, can now be computed from the $P(Y_{\mathrm{r}}(t), X_{\mathrm{u}}(t) \leqslant Y_{\mathrm{u}}(t)|k)$ values using Eq. (22). $\widehat{X}_{\mathrm{u}}(t)$ can subsequently be estimated using Eq. (21).

We note finally that in order to accommodate most completely the assumption of diagonal covariance matrices used in the estimation of cluster a posteriori probabilities, we initially estimate the distribution parameters of all clusters assuming diagonal covariance matrices in the implementation of the EM algorithm. We then compute full covariance matrices for all the clusters in a final pass of the algorithm.

6. Identifying unreliable components of the spectrogram

For missing-feature methods to be practicable, unreliable spectrographic components must be identified without a priori knowledge of their SNR. The two main approaches to this are based on computational auditory scene analysis (CASA) (e.g. Cooke et al., 1994a,b), and on explicit noise tracking (e.g. Drygajlo and El-Maliki, 1998; Vizhinho et al., 1999). CASA-based methods attempt to identify the reliable regions of the spectrogram based on acoustic cues and the known behavior of acoustic signals (e.g. grouping of spectral bands, harmonicity, etc.). Noise-tracking-based methods, on the other hand, attempt to maintain a running estimate of the noise spectrum and use this to determine which components of the spectrogram are unreliable.

In this paper we chose to use a Bayesian classifier to identify noisy components of the spectrogram. This reduces the task of identifying unreliable spectrographic components to a simple binary classification procedure. The features used in classification are 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. 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 (Seltzer et al., 2004). Fig. 4a shows the spectrogram from Fig. 1b, when unreliable components in the spectrogram have been identified from their known SNR values and removed. Fig. 4b shows the same figure when the identity of the unreliable components has been estimated by the Bayesian classifier used in this paper.

7. Experimental evalution

In this section we describe a series of experiments conducted to evaluate the recognition accuracy obtained using the proposed feature-compensation methods, and to contrast this with the accuracy obtained using state-based imputation and marginalization. Experiments were conducted on speech corrupted by white noise and segments of music. These noise types represent two extremes of spectral and temporal distortions—white noise has a flat spectrum and is stationary, while music has a very detailed spectral structure and is highly nonstationary. We initially describe experiments with "oracle" (or perfect) knowledge of the local SNR

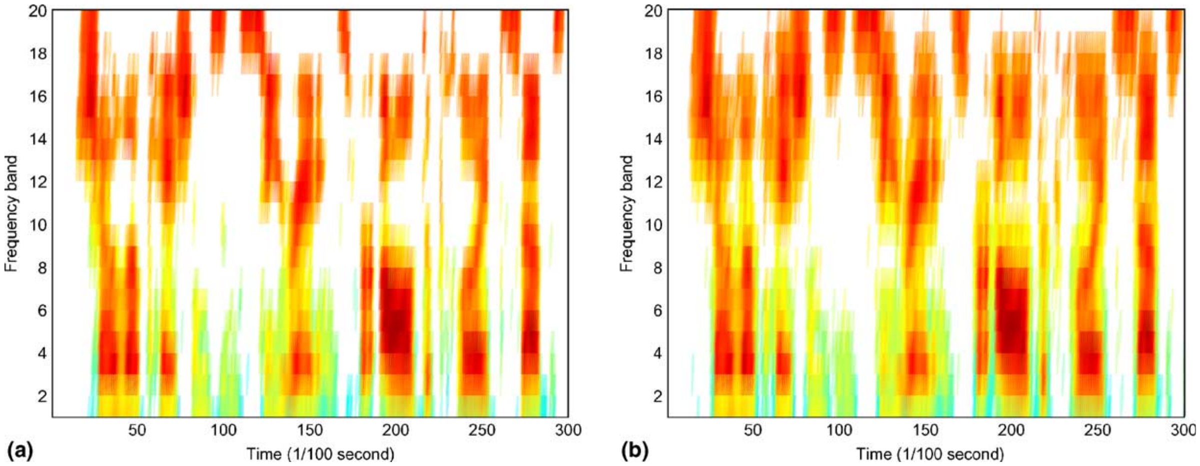


Fig. 4. (a) Mel spectrogram of the signal in Fig. 1b, when all components with SNR less than −5dB have been tagged unreliable. The white regions of the figure represent unreliable components. (b) Mel spectrogram for the same signal when the identity of unreliable regions has been estimated. The white regions in the figure represent components that have been identified by a classifier as being unreliable.

of time-frequency components in the spectrogram. Within these experiments we evaluate the effect of preprocessing and recognition with cepstra. These experiments establish an upper bound on the recognition performance obtainable with our experimental setup. We then describe results obtained from experiments employing a more realistic scenario where the locations of unreliable components must be estimated.

7.1. Experimental setup

The DARPA Resource Management (RM1) database (Price et al., 1988) and the CMU SPHINX-III HMM-based speech recognition system were used in all the experiments described in this paper. Context-dependent HMMs with 2000 tied states were trained using both the log spectra and cepstra of clean speech. State output distributions were modelled as Gaussian, except for the experiments that evaluated the performance of marginalization with more detailed state output distributions. In the latter case, state output distributions were modelled as mixtures of Gaussians. In all cases, the Gaussians in the state output distributions were assumed to have diagonal covariance matrices. A simple bigram language model was used. The language weight was kept to a minimum in all cases in order to emphasize the effect of the noisy acoustics on recognition accuracy. A 20-dimensional Mel spectral spectrographic representation was used in the experiments. Test utterances were corrupted by 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 component in the spectrograms of the noisy utterances.

7.2. Recognition performance with knowledge of true SNR

Missing feature methods depend critically on being able to identify unreliable regions of the spectrogram as such. In the experiments described in this section, we assume that this information is available and accurate. The recognition performance obtained with the various missing feature methods in this scenario represents an upper bound on the performance that can be obtained within the current experimental setup. Unreliable components of spectrograms were identified based on the true value of the SNR of time-frequency components, the computation of which was permitted by the experimental setup as explained in Section 7.1. All components whose SNR values lay below a threshold were deemed to be unreliable. A threshold value of 0 dB was found to be optimal or close to optimal at all SNRs for marginalization. For state-based imputation and the proposed feature-compensation methods, the best threshold across all noise levels was found to be -5 dB. The experiments reported in this section used these threshold values to identify unreliable components. For state-based imputation and marginalization, recognition was performed with the resulting incomplete spectrograms. For the feature-compensation methods, complete spectrograms were reconstructed. Fig. 5a and 5b show example spectrograms obtained by reconstructing unreliable components that have been estimated from their known SNR values, using correlation-based and cluster-based reconstruction. Recognition was performed using as features either the log-spectral vectors from the reconstructed spectrogram, or 13-dimensional cepstral coefficients derived from the log-spectral vectors.

7.2.1. Recognition with log spectra

Fig. 6 shows recognition accuracies obtained by applying the various missing-feature methods to speech corrupted by white noise and music to various SNRs. Recognition has been performed using log-spectral vectors in all cases. For marginalization, no mean normalization was performed on the features. For all other methods mean normalization was performed. In all cases, HMM state output distributions were modelled as Gaussian. We observe from these plots that marginalization is capable of resulting in remarkable robustness to corruption by noise. In fact, the recognition accuracy at 0 dB is only a relative 20% worse than that obtained at 25 dB. All other methods provide significant improvements over baseline recognition performance (with noisy vectors), but are much

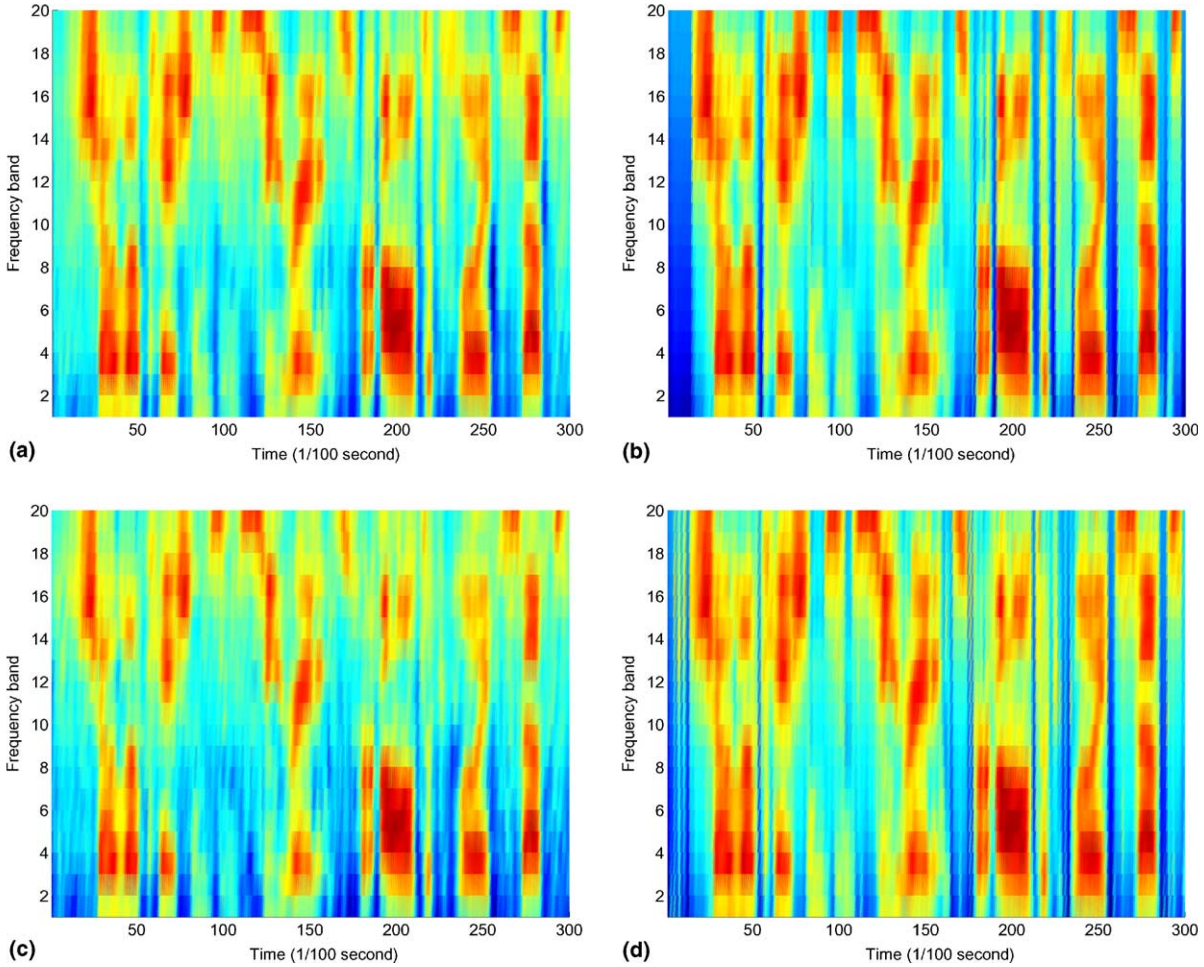


Fig. 5. Reconstruction of the Mel spectrogram in Fig. 4a. (a) Reconstruction obtained with correlation-based reconstruction when unreliable components have been identified based on their SNR. (b) Reconstruction obtained with cluster-based reconstruction when unreliable components have been identified based on their SNR. (c) Reconstruction obtained with correlation-based reconstruction when the identities of unreliable components have been estimated. (d) Reconstruction obtained with cluster-based reconstruction when the identities of unreliable components have been estimated.

worse that marginalization. This is to be expected when recognition is performed with log spectra, since, as mentioned in Section 4.2, marginalization performs *optimal* classification with the unreliable data, whereas the other methods do not. Feature-compensation methods do, however, perform comparably to, or better than state-based imputation.

7.2.2. *Effect of preprocessing*

The effect of preprocessing the signal is different on different missing feature methods. One form of preprocessing commonly used is mean normalization. In this procedure the mean value of the feature vectors is subtracted from all the vectors. This is known to result in significant improvement in recognition performance. When missing-feature methods are applied however, it is not clear whether this procedure is useful. Fig. 7a shows the effect of mean normalization on the recognition accuracy obtained with various missing-feature methods on speech corrupted to 10 dB by white noise. Both reliable and unreliable components

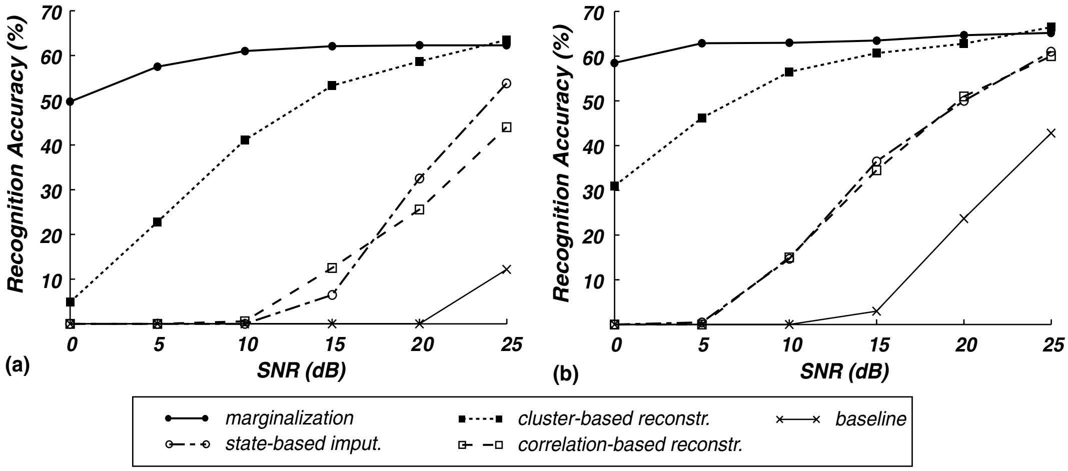


Fig. 6. Recognition performance of various missing feature methods on noisy speech, when unreliable components are located on the basis of their SNR values: (a) speech corrupted by white noise; (b) speech corrupted by music. In both figures the baseline recognition performance with the uncompensated noisy speech is also shown.

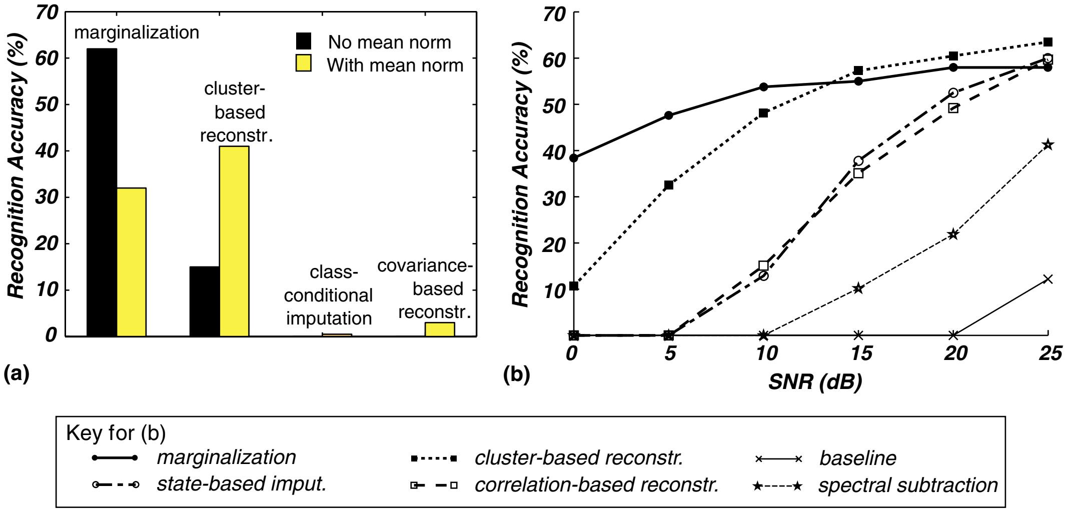


Fig. 7. Effect of preprocessing. (a) Comparison of recognition accuracies obtained with and without mean normalization of spectral vectors. (b) Recognition performance of various missing feature methods on speech corrupted by white noise to various SNRs, when spectral subtraction has been performed on the reliable portions of spectral vectors.

were used in computing the mean value of the vectors in all cases. We observe that mean normalization is useful in all cases where estimation of unreliable components is performed, i.e. for the feature-compensation methods and state-based imputation. For marginalization however, mean normalization actually results in a degradation of performance.

A basic assumption in missing feature methods is that the reliable components of noisy spectral vectors are good approximations to corresponding components of the underlying true vector. This, however, is not necessarily true, since the components that are identified as reliable can have fairly low SNR values, depending on the SNR threshold used to identify reliable components. When the corrupting noise is stationary or slowly varying, such as white noise, automobile noise, or factory noise, the spectrum of the noise can be reasonably well estimated and the SNR of the reliable components can be improved by performing spectral subtraction (Boll, 1979) as a preprocessing step. Fig. 7b shows the recognition performance obtained by the various methods on speech corrupted by white noise, when reliable spectral components have been preprocessed by spectral subtraction. As expected, spectral subtraction improves the performance of feature-compensation methods, as well as state-based imputation. However, it degrades the recognition performance of marginalization.

7.2.3. Recognition with cepstra

One of the primary arguments for spectrogram reconstruction methods is that the reconstructed spectrograms can now be used to derive cepstral features, and recognition can be performed with cepstra to obtain superior recognition performance. Fig. 8 shows the recognition results obtained with such a setup. Recognition with cepstra is greatly superior to that with log spectra. Comparison with Fig. 6 also shows that, although marginalization greatly outperforms other methods when recognition is performed with log spectra, the recognition performance obtained with cepstra derived from the reconstructed spectrograms results in much better recognition than obtainable with marginalization.

7.3. Effect of errors in identifying unreliable components

The experiments in the previous section only served to establish the upper bound performance obtainable for the various methods when location of unreliable components in the spectrogram is known *a priori*. In reality however, the location of unreliable components must be estimated. The estimation of these locations can be very errorful, and different missing-feature methods have differing sensitivity to errors in identifying unreliable components. Fig. 5c and d show the reconstructed spectrograms obtained for the spectrogram in Fig. 4b, where the identity of unreliable components was estimated. These figures are seen to be different from those in Fig. 5a and b obtained with *a priori* knowledge of the unreliable components.

Fig. 9 shows recognition accuracies obtained for several missing-feature methods applied to speech corrupted by white noise to 10 dB. Recognition has been performed using log spectra in all cases. We compare recognition accuracy obtained using perfect "oracle" knowledge of the true SNR values of spectrographic components to identify unreliable feature locations with the corresponding

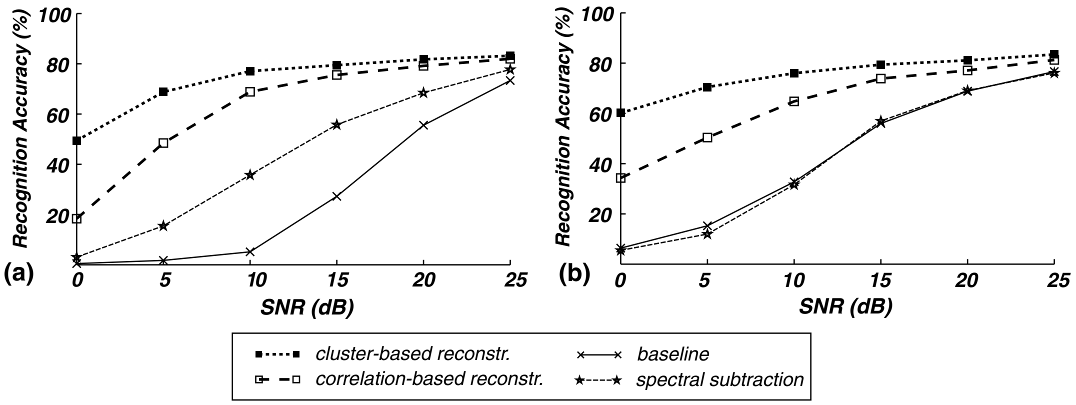


Fig. 8. Recognition performance obtained with cepstra derived from spectrograms reconstructed with prior knowledge of the identity of unreliable components. (a) Recognition on speech corrupted by white noise to various SNRs. (b) Recognition on speech corrupted by music to various SNRs. In both cases the baseline performance with uncompensated noisy speech, and the performance with a typical noise compensation algorithm, spectral subtraction, are shown for contrast.

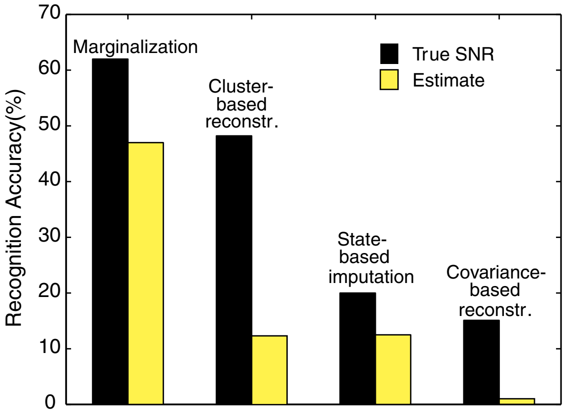


Fig. 9. Comparisons of recognition accuracy obtained when unreliable components are identified based on knowledge of their true SNR with accuracy obtained when the positions of unreliable components are estimated.

accuracy obtained when the locations of unreliable components are estimated from noisy data. Marginalization shows the greatest robustness to errors in estimation of unreliable components. In general, the classifier-compensation methods are much more robust to errors than the feature-compensation methods.

More detailed results are shown in Fig. 10, which shows recognition accuracy obtained with various missing-feature methods as a function of SNR on speech corrupted by white noise and music, when the identity of the unreliable components is estimated. In all cases, the HMM state output distributions were modelled by single Gaussians. Mean normalization was performed in the case of the feature-compensation methods and state-based imputation, but not for marginalization. Both classifier-compensation methods, marginalization and state-based imputation, are seen to outperform the feature-compensation methods. Marginalization, especially, is significantly superior to all other methods. The difference between marginalization and the other methods is further enhanced by its greater robustness to errors in identifying unreliable components.

Once again, however, reconstructed spectrograms can be used to derive cepstra for recognition. Fig. 11 shows the recognition performance obtained on speech corrupted by white noise and music with cepstra derived from spectrograms reconstructed by the proposed feature-compensation methods. Comparison with Fig. 10 reveals that even when the identities of unreliable components must be estimated, the recognition accuracy obtained with cepstra derived from reconstructed spectrograms is greater than that obtained with marginalization and log-spectra-based recognition.

In all experiments reported so far, state output distributions have been modelled by single Gaussians with diagonal covariances. It is likely that the recognition performance of the classifier-

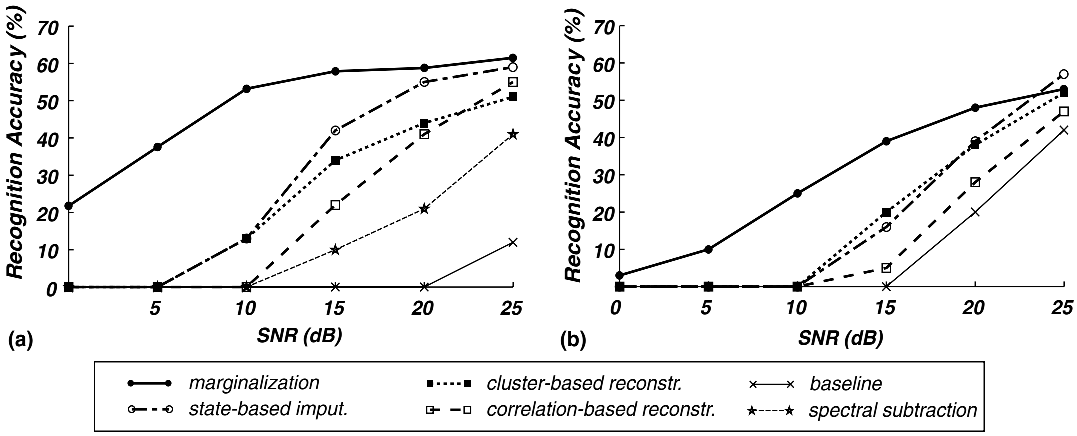


Fig. 10. Recognition performance of various missing feature methods on noisy speech when the locations of unreliable components is estimated: (a) speech corrupted by white noise; (b) speech corrupted by music. Recognition was performed using log spectra.

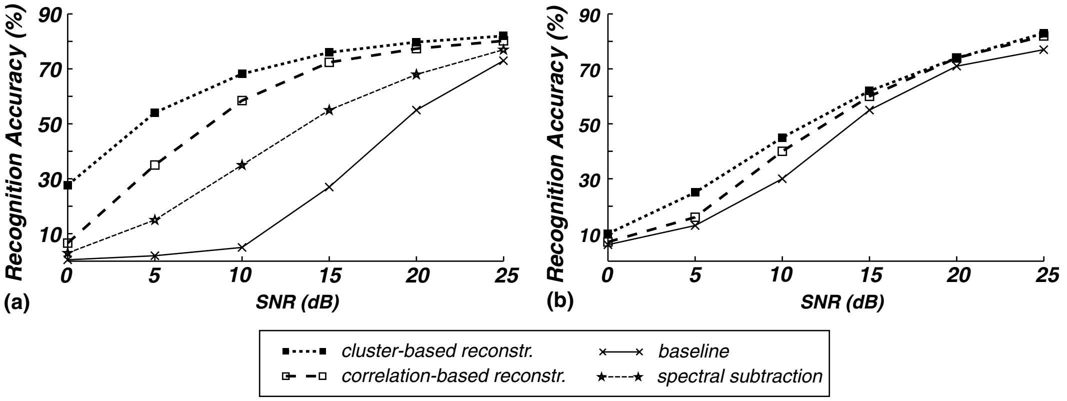


Fig. 11. Recognition with cepstra derived from reconstructed spectrograms, when the identity of unreliable components is estimated: (a) speech corrupted by white noise; (b) speech corrupted by music. As a contrast, baseline performance with the cepstra of noisy speech, and in the case of the white noise, performance with spectral subtraction are also shown.

compensation methods may be improved by modelling state output distributions by mixtures of Gaussians instead, thereby better capturing the correlations between spectral components. Fig. 12 tests this hypothesis. It shows the recognition performance obtained with marginalization when state output distributions are modelled by mixtures of 1, 2, 4 and 8 Gaussians, for speech corrupted by white noise and music, when the identities of unreliable components are estimated. The figure also shows the performance obtained from cepstra derived from spectrograms reconstructed by cluster-based reconstruction. It is seen that although increasing the number of Gaussians results in slightly better performance at higher SNRs, it still remains inferior to that obtained with the cepstra derived from reconstructed spectrograms. While the small improvements in recognition performance resulting from increasing the number of Gaussians in the state output densities is explained by the size of the training corpus for the RM1 database and greater improvements can

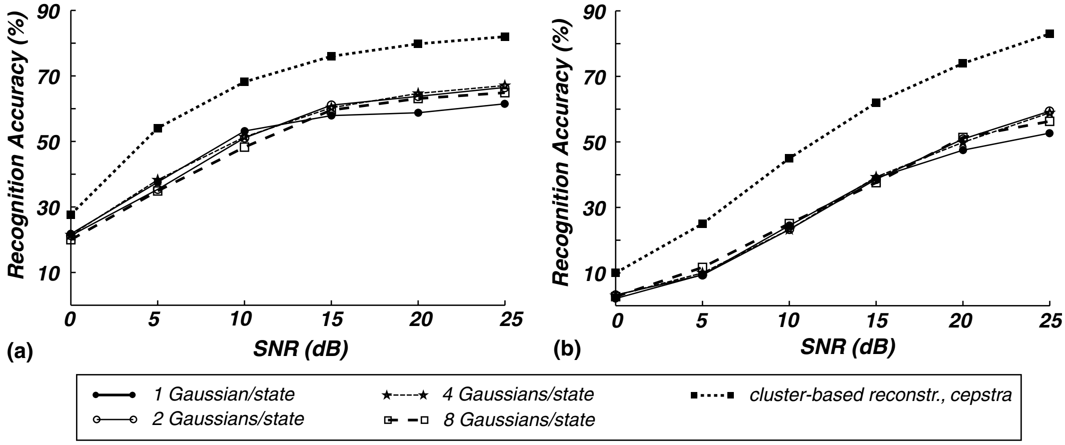


Fig. 12. Recognition performance of marginalization on HMMs with state output distributions modelled by mixtures of 1, 2, 4 and 8 Gaussians, when the identities of unreliable components are estimated: (a) recognition on speech corrupted by white noise; (b) recognition on speech corrupted by music. Recognition is performed with log spectra, in order to perform marginalization. The performance obtained with cepstra derived from spectrograms reconstructed by cluster-based reconstruction is also shown.

be expected with larger training corpora, we do not expect the trends in performance to change. In general, 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 components that are characteristic of marginalization. The advantage however diminishes as the SNR decreases to 0 dB or so.

7.4. Reconstructing spectrograms from HMM state sequences

In (Cooke et al., 1997) it has been suggested that classifier-compensation methods could be used to reconstruct spectrograms. One could, for instance, derive the best state-sequence for the utterance, and reconstruct the unreliable components of the spectral vectors using the distributions of the states with which they are associated. The reconstructed vectors could now be converted to cepstra for recognition. Fig. 13 shows the recognition accuracy obtained with cepstra derived from log spectra reconstructed in this manner, when state sequences were obtained using state-based imputation and marginalization. We note that overall, these methods are not more effective than the proposed feature-compensation methods.

In this experiment, the structure of the recognizer used to reconstruct the unreliable components was as complex as that used for the final recognition, i.e. both had as the same number of tied states (2000), and modelled state output densities as Gaussians. In principle, however, the HMMs used for the reconstruction can be much simpler than those used for recognition. While we have not explored this aspect, we point out that cluster-based reconstruction may be viewed as a limiting case where all states in the HMM used for reconstruction share a single Gaussian mixture distribution.

7.5. Computational complexity

The computational complexity of the various missing-feature methods also varies. Fig. 14 shows the average time in seconds taken by a 400-MHz DEC Alpha to recognize an utterance of speech from the RM database that has been 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 components, transformation to cepstra in the case of the feature-compensation methods, and recognition. The time taken for identifying unreliable components 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.

The differences in the computational requirements of the various methods is related to the mathematical operations underlying them. State-based

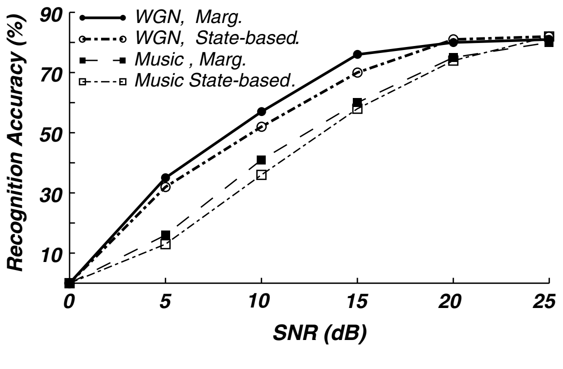


Fig. 13. Recognition accuracy using cepstra derived from log spectra reconstructed using state sequences hypothesized by classifier-compensation methods. Results are shown for speech corrupted both by white noise and music.

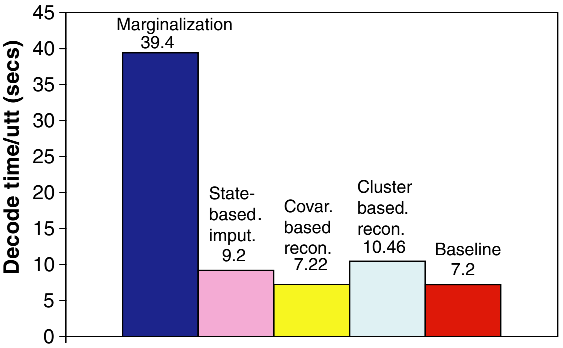


Fig. 14. Average time in seconds needed to recognize an utterance using different missing-feature methods.

imputation and covariance-based reconstruction only require MAP estimation of unreliable components and are relatively inexpensive. On the other hand, marginalization requires the computation of an error function for every unreliable component of a vector for every Gaussian in every HMM state whose output probability is evaluated. Cluster-based reconstruction similarly requires computation of error functions for every unreliable component for every cluster in the cluster-based representation. The number of Gaussians in the state output distributions of the HMMs in the recognizer is generally larger than that in the Gaussian mixture densities employed in cluster-based reconstruction. As a result, marginalization is computationally more expensive than cluster-based reconstruction.

Since the computational expense of marginalization (and that of classifier-compensation methods in general) is a function of the number of HMM states for which output probabilities must be evaluated, it is also related to the perplexity of the language model used by the recognizer. The number of active hypotheses considered by the recognizer at any instant increases with the perplexity of the language model used. This in turn increases the number of HMM states that must be evaluated for each frame, and hence the number of error functions that must be computed. The computational complexity of all missing-feature methods is also related to the SNR of the data. The number of corrupted spectral components that must be marginalized or reconstructed increases with decreasing SNR. Decreasing SNR also has a secondary effect on classifier-compensation methods: the number of active hypotheses considered by the recognizer that survive pruning, and hence the number of HMM states to be evaluated, usually increases with decreasing SNR.

On the other hand, the actual computation required by the various methods is also dependent on the manner in which they have been implemented. For instance, the computation of error functions can be considerably speeded up by the use of lookup tables. This would speed up both marginalization and cluster-based reconstruction, the former more than the latter. Hence, while the comparisons shown in Fig. 14 may be considered indicative of the relative complexity of the various methods, the actual computational complexity of the methods would vary with the recognition task and the specific implementation of the algorithms.

8. Conclusions

While the actual recognition results shown in Section 7 are specific to a particular database, experimental setup, and recognition system used, they establish a set of very consistent trends. Of all the missing feature methods, marginalization is clearly the best when recognition is performed with log spectral vectors. It results in the most robustness to noise and errors in identifying unreliable components. It must be emphasized that when recognition is performed in the feature domain where unreliable components are identified (i.e. on spectra or log spectra), the best classifier-compensation methods can always be expected to outperform the best feature-compensation methods. In addition, in classifier-compensation methods the search algorithm used by the recognizer can itself be modified to account for the uncertainty in the location of corrupt spectrographic components (e.g. Barker et al., 2003).

The proposed feature-compensation methods are observed to result in better eventual recognition performance than marginalization primarily because they permit recognition with cepstra derived from the reconstructed spectrograms. Of the two methods proposed, cluster-based reconstruction provides significantly better accuracy than correlation-based reconstruction. The latter algorithm, however, has the advantages that it is extremely simple, and that it provides better performance than state-based imputation when recognition is performed using cepstra. In addition, in other experiments not reported in this paper correlation-based reconstruction was found to be superior to cluster-based reconstruction when the loss of spectrographic information was due to the random excision of time-frequency components, e.g. by loss during transmission, and not due to additive noise. The feature-compensation methods described in this paper use only very simple statistical models to represent the distribution of the spectral vectors of clean speech. It is expected that their performance can be improved by using more sophisticated models. Cluster-based reconstruction is expected to gain by adding temporal dependencies in the statistical model, either by modelling the *a priori* probabilities using a Markov chain, effectively converting the Gaussian mixture distribution to an HMM, or by modelling the distribution of temporal derivatives of the vectors jointly with the vectors, or by some combination of the two. Similarly covariance-based reconstruction may be improved by modelling spectrograms as the output of a mixture of stochastic processes.

Renevey and Drygajlo (2000) and Morris et al. (2001) have shown that the performance of marginalization may be improved significantly by associating a probability of reliability with spectrographic components, rather than by tagging them as reliable or not in a binary manner. Such probabilistic tagging can also be incorporated into the methods proposed in this paper.

Finally, it must be pointed out that the feature-compensation methods are not limited to working only with the statistical models used in this paper. Since the basic idea behind these methods is to reconstruct the spectrograms externally to the recognizer, other techniques, such as Kalman filters or neural networks might also be used to reconstruct the unreliable components.

Acknowledgments

This research was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this publication does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

Appendix A. Iterative procedure for bounded MAP estimation

The problem of joint bounded MAP estimation is to find a set of values $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k$ such that

$$\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_k = \arg\max_{x_1, x_2, \ldots, x_k} \{P(x_1, x_2, \ldots, x_k | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k)\} \tag{A.1}$$

We derive an iterative solution for this estimate in this appendix.

Let $x_1^n, x_2^n, \ldots, x_k^n$ be the estimate obtained after the nth iteration of this procedure. If the $n + 1$th estimate of x_1 is obtained as

$$x_1^{n+1} = \arg\max_{x_1} \{P(x_1, x_2^n, \ldots, x_k^n | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k)\} \tag{A.2}$$

then it is easy to see that

$$P(x_1^{n+1}, x_2^n, \ldots, x_k^n | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k) \geqslant P(x_1^n, x_2^n, \ldots, x_k^n | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k) \tag{A.3}$$

Using Bayes' rule and eliminating all irrelevant terms, Eq. (A.2) can be restated as

$$x_1^{n+1} = \arg\max_{x_1} \{P(x_1 | x_1 \leqslant Y_1, x_2^n, \ldots, x_k^n)\} \tag{A.4}$$

which is simply the bounded MAP estimate of x_1, conditioned on $x_2^n, \ldots, x_k^n$. When $P(x_1, x_2^n, \ldots, x_k^n)$ is Gaussian, this is simply given by

$$x_1^{n+1} = \min(Y_1, E[x_1 | x_2^n, \ldots, x_k^n]) \tag{A.5}$$

It can similarly be shown that if the $n + 1$th estimate of x_j is obtained as

$$x_j^{n+1} = \arg\max_{x_1} \{P(x_j | x_1^{n+1}, x_2^{n+1}, \ldots, x_{j-1}^{n+1}, x_j \leqslant Y_j, x_{j+1}^n, \ldots, x_k^n)\} \tag{A.6}$$

then

$$P(x_1^{n+1}, x_2^{n+1}, \ldots, x_{j-1}^{n+1}, x_j^{n+1}, x_{j+1}^n, \ldots, x_k^n | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k) \geqslant P(x_1^{n+1}, x_2^{n+1}, \ldots, x_{j-1}^{n+1}, x_j^n, x_{j+1}^n, \ldots, x_k^n | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k) \tag{A.7}$$

In other words, if we begin with some set of initial estimates $x_1^1, x_2^1, \ldots, x_k^1$, and iteratively find the $n + 1$th estimate of each x_j as the bounded MAP estimate of that component as given by Eq. (A.6), $P(x_1, x_2, \ldots, x_k | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k)$ is guaranteed not to decrease at each step in the iteration.

For Gaussian random variables, Eq. (A.6) can be equivalently written as

$$x_j^{n+1} = \min(Y_j, E[x_j | x_1^{n+1}, x_2^{n+1}, \ldots, x_{j-1}^{n+1}, x_{j+1}^n, \ldots, x_k^n]) \quad (A.8)$$

When $P(x_1, x_2, \ldots, x_k)$ is Gaussian, $P(x_1, x_2, \ldots, x_k | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k)$ has only one peak. Thus, the iterative solution given by Eq. (A.8) is guaranteed to find this peak, which is the unique solution to Eq. (A.1). Therefore, the iterative solution to the joint bounded MAP estimation of a set of jointly Gaussian variables $x_1, x_2, \ldots, x_k$ conditioned on the bound $x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k$ is given by the following procedure:

(1) Initialize all the x_i values as $x_i^1 = Y_i$
(2) Obtain the $n+1$th estimate of x_j as
$$x_j^{n+1} = \min(Y_j, E[x_j | x_1^{n+1}, x_2^{n+1}, \ldots, x_{j-1}^{n+1}, x_{j+1}^n, \ldots, x_k^n])$$
(3) Iterate until $P(x_1, x_2, \ldots, x_k | x_1 \leqslant Y_1, x_2 \leqslant Y_2, \ldots, x_k \leqslant Y_k)$ converges.

References

Acero, A., 1993. Acoustic and Environmental Robustness in Automatic Speech Recognition. Kluwer Academic Publishers, Boston, MA.

Barker, J., Cooke, M., Ellis, D.P.W.E., 2003. Decoding speech in the presence of other sources. Elsewhere in this issue.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. 27, 113–120.

Cooke, M.P., Green, P.G., Crawford, M.D., 1994. Handling missing data in speech recognition. In: Proc. Internat. Conf. on Speech and Language Processing, pp. 1555–1558.

Cooke, M., Green, P., Anderson, C., Abberley, D., 1994. Recognition of occluded speech by Hidden Markov Models. Technical Report TR-94-05-01, Department of Computer Science, University of Sheffield, 1994.

Cooke, M.P., Morris, A., Green, P.D., 1997. Missing data techniques for robust speech recognition. In: Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing, Munich, Germany.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and uncertain acoustic data. Speech Commun. 34, 267–285.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28, 357–366.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B 39, 1–38.

Drygajlo, A., El-Maliki, M., 1998. Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. In: Proc. ICASSP‘98, Seattle, WA, pp. 121–124.

Dupont, S., 1998. Missing data reconstruction for robust automatic speech recognition in the framework of hybrid HMM/ANN systems. In: Proc. ICSLP98, Sydney, Australia, December 1998.

Fletcher, H., 1953. Speech and Hearing in Communication. Van Nostrand, New York.

Gales, M.J.F., Young, S.J., 1996. Robust continuous speech recognition using parallel model combination. IEEE Trans. Speech Audio Process. 4, 352–359.

Josifovski, L., Cooke, M., Green, P., Vizihno, A., 1999. State based imputation of missing data for robust speech recognition and speech enhancement. In: Proc. EUROSPEECH 99, Budapest, Hungary.

Linde, Y., Buzo, A., Gray, R.M., 1980. An algorithm for vector quantizer design. IEEE Trans. Commun. COM-28, 84–95.

Lippmann, R., Carlson, B., 1997. Using missing feature theory actively select features for robust speech recognition with interruptions, filtering, and noise. In: Proc. Eurospeech97, Rhodes, Greece, pp. 37–40.

McQueen, J., 1967. Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symp. on Mathematics, Statistics and Probability, vol. 1, pp. 281–298.

Miller, G.A., Licklider, J.C.R., 1950. The intelligibility of interrupted speech. J. Acoustic Soc. Am. 22, 167–173.

Moreno, P.J., 1996. Speech recognition in noisy environments. Ph.D. Dissertation, Carnegie Mellon University.

Morris, A.C., Barker, J., Bourlard, H., 2001. From Missing Data to Maybe Useful Data: Soft Data Modelling for Noise Robust ASR. WISP 2001, Stratford-upon-Avon, UK.

O'Shaughnessy, D., 1987. Speech Communication—Human and Machine. Addison-Wesley.

Papoulis, A., 1991. Probability, random variables, and stochastic processes, third ed. McGraw-Hill, New York.

Price, P., Fisher, W.M., Bernstein, J., Pallet, D.S., 1988. The DARPA 1000 word Resource Management database for continuous speech recognition. In: Proc. IEEE Conf. on Acoustics Speech and Signal Processing, pp. 651–654.

Raj, B., 2000. Reconstruction of incomplete spectrograms for robust speech recognition. Ph.D. Thesis, Carnegie Mellon University.

Raj, B., Parikh, V., Stern, R.M., 1997. The effects of background music on speech recognition accuracy. In: Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, Munich, Germany.

Raj, B., Singh, R., Stern, R.M., 1998. Inference of missing spectrographic features for robust speech recognition. In: Proc. ICSLP98, Sydney, Australia.

Renevey, P., 2001. Speech in noisy conditions using missing feature approach. Ph.D. dissertation, Thése EPFL, No. 2303, Swiss Federal Institute of Technology.

Renevey, P., Drygajlo, A., 1999. Missing feature theory and probabilistic estimation of the clean components for robust speech recognition. In: Proc. Eurospeech99, Budapest, Hungary, pp. 2627–2630.

Renevey, P., Drygajlo, A., 2000. Introduction of a reliability measure in missing data approach for robust speech recognition. In: Proc. EUSIPCO 2000.

Seltzer, M.L., Raj, B., Stern, R.M., 2004. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Commun., this issue, doi: 10.1016/j.specom.2004.03.006.

Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, 1990, pp. 845–848.

Vizhinho, A., Green, P., Cooke, M., Josifovski, L., 1999. Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: an integrated study. In: Proc. Eurospeec99, Budapest, Hungary, pp. 2407–2410.

Warren, R.M., Riener, K.R., Bashford, J.A., Brubaker, B.S., 1995. Spectral redundancy: intelligibility of sentences heard through narrow spectral slits. Percept. Psychophys. 57, 175–182.