Bhiksha Raj and Richard M. Stern

Improving recognition accuracy in noise by using partial spectrographic information

Missing-Feature Approaches in Speech Recognition

Despite decades of focused research on the problem, the accuracy of automatic speech recognition (ASR) systems is still adversely affected by noise and other sources of acoustical variability. For example, there are presently dozens, if not hundreds, of algorithms that have been developed to cope with the effects of quasistationary additive noise and linear filtering of the channel [16], [23]. While these approaches are reasonably effective in the context of their intended purposes, they are generally ineffective in improving recognition accuracy in many more difficult environments. Some of these more challenging conditions, which are frequently encountered in some of the most important potential speech recognition applications today, include speech in the presence of transient and nonstationary disturbances (as in many factory, military, and telematic environments), speech in the presence of background music or background speech (as in the automatic transcription of broadcast program material as well as in many natural environments), and speech at very low signal-to-noise ratios (SNRs).

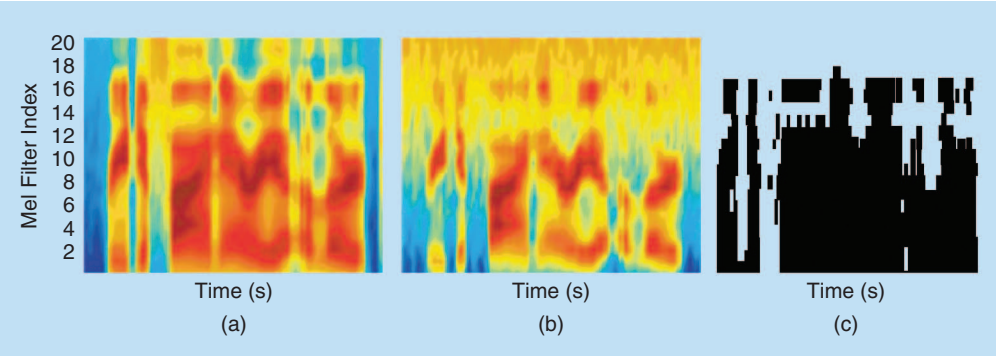
Conventional environmental compensation provides only limited benefit for these problems even today. For example, Raj et al. [29] showed that while codeword-dependent cepstral normalization is highly effective in dramatically reducing the impact of additive broadband noise on speech recognition accuracy, it is relatively ineffective when speech is presented to the system in the presence of background music, for reasons that are believed to be a consequence of the nonstationary nature of the acoustical degradation.

1053-5888/05/$20.00©2005IEEE

This article describes and discusses alternative robust recognition approaches that have come to be known as missing feature approaches [9], [10], [20]. These approaches are based on the observation that speech signals have a high degree of redundancy—human listeners are able to comprehend speech that has undergone considerable spectral excisions. For example, normal conversation is possible with speech that has been either high- or low-pass filtered with a cutoff frequency of 1,800 Hz [17].

Briefly, in missing feature approaches, one attempts to determine which cells of a spectrogram-like time-frequency display of speech information are unreliable (or missing) because of degradation due to noise or to other types of interference. The cells that are determined to be unreliable or missing are either ignored in subsequent processing and statistical analysis (although they may provide ancillary information), or they are filled in by optimal estimation of their putative values.

The application of missing feature techniques to robust automatic speech recognition has been strongly influenced by two complementary fields of signal analysis. Many of the mathematical approaches to missing feature recognition were first developed for the task of completing partially occluded objects in visual pattern recognition [1]. In addition, there is a strong interchange of techniques and applications between missing feature recognition and the field of computational auditory scene analysis (CASA), which seeks to segregate, identify, and recognize the separate components of a complex sound field [8]. The mathematics developed for missing feature recognition can be applied to other types of signal restoration implicit in CASA, and the analysis approaches developed in CASA can be useful for identifying degraded components of signals that can potentially be restored through the use of missing feature techniques.



[FIG1] (a) The "mel" spectrogram for an utterance of clean speech. (b) The mel spectrogram for the same utterance when it has been corrupted by white noise to an SNR of 10 dB. (c) Spectrographic mask for the noise-corrupted utterance using an SNR threshold of 0 dB.

SPECTRAL MEASUREMENTS AND SPECTROGRAPHIC MASKS

Missing feature methods model the effect of noise on speech as the corruption of regions of time-frequency representations of the speech signal. A variety of time-frequency representations may be used for the purpose. Most of them follow the same basic procedure: the speech signal is first segmented into overlapping frames, typically about 25 ms wide. From each frame, a power spectrum is estimated as

$$Z_p(m,k) = \left| \sum_{n=0}^{N-1} w(n-mL)x(n)e^{-j2\pi(n-mL)k/N} \right|^2, \qquad (1)$$

where $Z_p(m,k)$ represents the kth frequency band of the power spectrum of the mth frame of speech, $w(n)$ represents a window function, $x(n)$ represents the nth sample of the speech signal, L represents the shift in samples between adjacent frames, and N represents the number of samples within the segment. The power spectral components are often reduced to a smaller number of components using triangular weighting functions that represent the frequency responses of the filters in a filter bank that is designed to mimic the frequency sensitivity of the human ear

$$X_p(m,k) = \sum_j h_k(j) Z_p(m,j), \qquad (2)$$

where $h_k(j)$ represents the frequency response of the kth filter in the filter bank. The most commonly used filter banks are the Mel filter bank [13] and the ERB filter bank [21]. Alternately, $X_p(m,k)$ may be obtained by sampling the envelope of signals at the output of a filter bank [25]. We will refer to both the power spectra given by (1) and the combined power spectra of (2) generically as power spectra.

The power spectral components are compressed by a nonlinear function to obtain the final time-frequency representation, $X(m,k) = f(X_p(m,k))$, where $f(\cdot)$ is usually a logarithm or a cube root. We assume, without loss of generality, that $f(\cdot)$ is a logarithm. The outcome of the analysis is a two-dimensional representation of the speech signal, which we will refer to as a *spectrogram*. Figure 1(a) shows a pictorial representation of the spectrogram for a clean speech signal.

When the speech signal is corrupted by uncorrelated additive noise, the power spectrum of the resulting signal is the sum of the power spectra of the clean speech and the noise

$$Y_p(m,k) = X_p(m,k) + N_p(m,k), \qquad (3)$$

where $Y_p(m,k)$ and $N_p(m,k)$ represent the kth frequency bands of the power spectrum of the noisy speech and the noise, respectively, in the mth frame of the signal. Since all power

spectral terms are nonnegative, $Y_p(m,k)$ is nearly always greater than or equal to $X_p(m,k)$. (In reality, finite-sized samples of uncorrelated processes are rarely perfectly orthogonal, and the power spectra of two uncorrelated processes do not simply add within any particular analysis frame. To account for this, an additional factor of $2\sqrt{X_p(m,k)N_p(m,k)}\cos(\theta)$ must be included in (3), where θ is the angle between the kth term of the complex spectra of the speech and the noise. In practice, however, this term is usually small and can safely be ignored.) The SNR of $Y_p(m,k)$, the (m,k)th component of the spectrogram of the noisy signal, is given by $X_p(m,k)/N_p(m,k)$. The SNR of the spectrogram varies with both time and frequency. Typically, for any level of noise, the spectrogram includes regions of very high SNR (which are dominated by contributions of the speech component), as well as regions of very low SNR (which represent the characteristics of the noise more than the underlying speech). The low SNR regions of the spectrogram adversely affect speech recognition accuracy. As the overall SNR of a noisy utterance decreases, the proportion of low SNR regions increases, resulting in worse recognition accuracy.

In missing feature methods, it is assumed that all components of a spectrogram with SNR above some threshold T are reliable estimates of the corresponding components of clean speech. In other words, it is assumed that the observed value $Y(m,k)$ of such components is equal to the value $X(m,k)$ that would have been obtained had there been no noise. Spectral components $Y(m,k)$ with SNRs below the threshold, however, are assumed to be unreliable and assumed not to represent $X(m,k)$. They merely provide an upper bound on the true value of $X(m,k)$, *i.e.,* $X(m,k) \leq Y(m,k)$. The set of tags that identify reliable and unreliable components of the spectrogram are referred to as the *spectrographic mask* for the utterance. Recognition must now be performed with the resulting incomplete measurements of the speech signal, where only some of the spectral components are reliably known and the rest are essentially unknown. Figure 1(b) shows the spectrogram for the utterance in Figure 1(a) after it has been corrupted to 10 dB by white noise. Figure 1(c) shows the spectrographic mask for the utterance when a threshold of 0 dB is used to identify unreliable elements, with the black regions representing the values of m and k, for which the corresponding spectral values $Z_p(m,k)$ are to be considered reliable. While in practice there is no single optimal threshold [28], [33], Cooke et al. and Raj et al. have typically used thresholds that lie between 5 and -5 dB.

Before we proceed, we must establish some of the notation and terminology used in the rest of the article. We represent the observed power spectrum for the mth frame of speech as a vector $Y_p(m)$ and the individual frequency components of $Y_p(m)$ as $Y_p(m,k)$. We assume that the final time-frequency representation consists of sequences of log-spectral vectors derived as the logarithm of the power spectrum. We represent the mth log spectral vector as $Y(m)$ and the individual frequency components in it as $Y(m,k)$. For noisy speech signals, the observed signal is assumed to be a combination of a clean speech signal and noise. We represent the power spectrum of the clean speech in the mth frame of speech as the vector $X_p(m)$, the corresponding log-spectral vector as $X(m)$, and the individual frequency components of the two as $X_p(m,k)$ and $X(m,k)$, respectively. In addition, we represent the power spectrum of the corresponding noise and its frequency components as $N_p(m)$ and $N_p(m,k)$. We represent the sequence of all spectral vectors $Y(m)$ for any utterance as $\mathbf{Y}$ and the sequence of all corresponding $X(m)$ as $\mathbf{X}$. For brevity, we refer to $Y(m)$, $X(m)$, and $N(m)$ as spectral vectors (instead of log-spectral vectors) and to $\mathbf{Y}$ and $\mathbf{X}$ as spectrograms. We refer to the time and frequency indices (m,k) as a *time-frequency location* in the spectrogram and the corresponding components $X(m,k)$ and $Y(m,k)$ as the spectral components at that location.

Ideally, recognition would be performed with $\mathbf{X}$, the spectrogram for the clean speech (or with features derived from it). Unfortunately, the spectrogram $\mathbf{X}$ is obscured by the corrupting noise, resulting in the observation of a noisy spectrogram $\mathbf{Y}$ which differs from $\mathbf{X}$. Every vector $Y(m)$ in $\mathbf{Y}$ has a set of reliable components that are good approximations to the corresponding components of $X(m)$. We represent the spectrographic mask that distinguishes reliable and unreliable components of a spectrogram by $\mathbf{S}$. The elements of $\mathbf{S}$ may be either binary tags identifying spectral components as reliable or not with certainty, or they may be real numbers between zero and one that represent a measure of confidence in the reliability of the spectral components. For the rest of this section, we assume that the elements of $\mathbf{S}$ are binary. The spectrographic mask vector that identifies unreliable and reliable components of a single spectral vector $Y(m)$ is represented as $S(m)$. We arrange the reliable components of $Y(m)$ into a vector $Y_r(m)$ and the corresponding components of $X(m)$ into a vector $X_r(m)$. Each vector $Y(m)$ also has a set of unreliable components that provide an upper bound on the value of the corresponding components of $X(m)$. We arrange the unreliable components of $Y(m)$ into a vector $Y_u(m)$, and the corresponding components of $X(m)$ into $X_u(m)$. We refer to the set of all $Y_r(m)$ as $\mathbf{Y_r}$. Similarly, we refer to the set of all $Y_u(m)$, $X_r(m)$, and $X_u(m)$ as $\mathbf{Y_u}$, $\mathbf{X_r}$, and $\mathbf{X_u}$, respectively. The relationship between $Y_r(m)$, $X_r(m)$, $Y_u(m)$, and $X_u(m)$ is given by

$$\begin{aligned} X_r(m) &= Y_r(m) \\ X_u(m) &\leq Y_u(m) \end{aligned} \tag{4}$$

It is also common to assume a lower bound on $X_u(m)$, based on a priori knowledge of typical feature values. For instance, when the nonlinear compressive function $f(\cdot)$ used in the computation of the spectrogram is a cube root, $X_u(m)$ is assuredly lower bounded at zero. In some sections of the article, we drop the time and frequency-component indices of vectors, simply representing them as X, X_r, and X_u, where these vector indices are not critical to comprehension, for brevity of notation. This should not cause any confusion.

ADDITIONAL BACKGROUND

In this section, we briefly describe three other topics that are of relevance to the rest of this article: speech recognition based on hidden Markov models (HMMs), bounded marginalization of Gaussian densities, and bounded maximum a posteriori estimation of Gaussian random variables.

SPEECH RECOGNITION USING HMMS

While the experimental results described in this article were obtained using HMMs, the concepts described are easily carried over to other types of recognition systems as well. The technology of HMMs has been described in detail in other sources [27], but we will summarize the basic computation briefly to introduce the notational conventions used in this article.

Given a sequence of feature vectors $\mathbf{X}$ derived from an utterance, ASR systems seek to identify $\hat{W}$, the sequence of words in that utterance per the optimal Bayesian classifier

$$\hat{W} = \arg\max_{W}\{P(W|\mathbf{X})\} = \arg\max_{W}\{P(\mathbf{X}|W)P(W)\}. \quad (5)$$

$P(W)$ is the a priori probability that the word sequence W was uttered and is usually specified by a *language model*. $P(\mathbf{X}|W)$ is the likelihood of $\mathbf{X}$, given that W was the sequence of words uttered.

The distribution of the feature vectors for W is modeled by an HMM that assumes that the process underlying the signal for W transitions through a sequence of states s from frame to frame. Each transition produces an observed vector that depends only on the current state. Let $P_W(X|s)$ denote the state output distribution of state s of the HMM for W. Ideally, $P(\mathbf{X}|W)$ must be computed considering every state sequence through the HMM. In practice, however, ASR systems attempt to estimate the best state sequence jointly with the best word sequence. In other words, recognition is performed as

$$\hat{W} = \arg\max_{W}\max_{s}\left\{P(w)P(s|W)\left(\prod_{m}P_W(X(m)|s)\right)\right\}, \quad (6)$$

where s represents the state sequence $s_1, s_2, \ldots$, and $P(s|W)$ is the probability of s as obtained from the transition probabilities of the HMM for W. The state output distribution terms $P_W(X|s)$ in (6) lie at the heart of how missing feature methods affect speech recognition.

BOUNDED MARGINALIZATION OF GAUSSIAN DENSITIES

Let a random vector X have a Gaussian density with mean μ and a diagonal covariance matrix Θ. Let the indices of the components of X be segregated into two mutually exclusive sets, U and R. Let X_u be a vector constructed from all components of X whose indices lie in U. Let X_r be a vector constructed from all components of X whose indices lie in R. Let it be known that X_u is bounded from above by H_u and below by L_u. It can be shown that the marginal probability density of X_r is given by

$$P(X_r, L_u \le X_u \le H_u; \mu, \Theta) = \prod_{j\in R}\frac{1}{\sqrt{2\pi\sigma(j)}}e^{\frac{-(X(j)-\mu(j))^2}{2\sigma(j)}} \times \prod_{l\in U}\int_{L(l)}^{H(l)}\frac{1}{\sqrt{2\pi\sigma(l)}}e^{\frac{-(x-\mu(l))^2}{2\sigma(l)}}dx, \quad (7)$$

where the arguments after the semicolon on the left hand side of the equation represent the parameters of the distribution. $X(j)$, $\mu(j)$, and $\sigma(j)$ represent the j th components of X and μ and the jth diagonal component of Θ, respectively. $H(l)$ and $L(l)$ represent the components of H_u and L_u that correspond to $Y(l)$. The integral term to the right marginalizes out X_u within its known bounds.

BOUNDED MAP ESTIMATION OF GAUSSIAN RANDOM VARIABLES

Let X be a random vector drawn from a Gaussian density with mean μ and covariance Θ. X is corrupted to generate an observed vector Y such that some components of Y are identical to the corresponding values of X, while the remaining components only provide an upper bound on the corresponding values of X. Let X_r and X_u be vectors constructed from the known and unknown components of X, and let Y_r and Y_u be constructed from the corresponding components of Y. The bounded MAP estimate of X_u is given by

$$\hat{X}_u = \arg\max_{X_u} P(X_u|X_r = Y_r, X_u \le Y_u). \quad (8)$$

Note that when Θ is a diagonal matrix, the bounded MAP estimate of X_u is simply $\min(\mu_u, Y_u)$, where μ_u is the expected value of X_u and is constructed from the corresponding components of μ. When Θ is not a diagonal matrix, an iterative procedure is required to obtain X_u as follows.

The unbounded MAP estimate of a log-spectral component $X(i)$, given that the values of all other components $X(k) = \bar{X}(k)$, $k \ne i$ is

$$\tilde{X}(i) = \arg\max_{X(i)} P(X(i)|X(k) = \bar{X}(k),\ k \ne i) = \mu(i) + \frac{1}{\sigma(i)}\Theta_{X(i),\bar{X}}(\bar{X} - \bar{\mu}), \quad (9)$$

where $\bar{X}$ is a vector constructed from $\bar{X}(k)$, $k \ne i$, $\bar{\mu}$ is its mean value, $\Theta_{X(i),\bar{X}}$ is a row matrix representing the cross covariance between $X(i)$ and $\bar{X}$, and $\mu(i)$ and $\sigma(i)$ are the mean and covariance of $X(i)$. $\bar{\mu}$, $\sigma_{X(i),\bar{X}}$, $\mu(i)$ and $\sigma(i)$ can be constructed from the components of μ and Θ. The iterative procedure for bounded MAP estimation is [28] the following:

1) Initialize $\bar{X}(k) = Y(k)\ \forall k$.
2) For each of $X(i)$, $i \in U$

$$\tilde{X}(k) = \arg\max_{X(k)} P(X(k)|X(i) = \bar{X}(i),\ i \neq k)$$
$$\bar{X}(k) = \min(\tilde{X}(k), Y(k)) \tag{10}$$

3) Iterate Step 2 until the $\bar{X}(k)$ have converged. The bounded MAP estimate $\hat{X}_u$ is constructed from the converged values of $\bar{X}(i), i \in U$.

In the following, we will represent the bounded MAP estimate described by (8) as BMAP$(X_u|X_r = Y_r, X_u \leq Y_u)$, or more explicitly, as BMAP$(X_u|X_r = Y_r, X_u \leq Y_u; \mu, \Theta)$.

RECOGNITION WITH UNRELIABLE SPECTROGRAMS

Using the notation developed previously, the problem of recognition with unreliable spectrograms can be stated as follows: we desire to obtain a sequence of feature vectors, $\mathbf{X}$, from the speech signal for an utterance and estimate the word sequence that it represents. Instead, we obtain a corrupted set of feature vectors $\mathbf{Y}$, with a reliable subset of components $\mathbf{Y_r}$ that are a close approximation to $\mathbf{X_r}$, i.e., $\mathbf{X_r} \approx \mathbf{Y_r}$, and an unreliable subset $\mathbf{Y_u}$, which merely provides an upper bound on $\mathbf{X_u}$, i.e., $\mathbf{X_u} \leq \mathbf{Y_u}$. We must now perform recognition with only this partial knowledge of $\mathbf{X}$.

There are two major approaches to solving this problem:

- *Feature-vector imputation:* estimate $\mathbf{X_u}$ to reconstruct a complete uncorrupted feature vector sequence $\mathbf{X_r} \cup \mathbf{X_u}$ and use it for recognition.
- *Classifier modification:* modify the classifier to perform recognition using $\mathbf{X_r}$ and the unreliable $\mathbf{Y_u}$ itself.

FEATURE-VECTOR IMPUTATION: ESTIMATING UNRELIABLE COMPONENTS

The damaged regions of spectrograms are reconstructed from the information available in the reliable regions and a priori knowledge about the structure of speech spectrograms that has been obtained from a training corpus of uncorrupted speech. We describe two reconstruction techniques: 1) cluster-based reconstruction, in which damaged components are reconstructed based solely on the relationships among the components within individual vectors, and 2) covariance-based reconstruction, in which reconstruction considers statistical correlations among all components of the spectrogram [28], [30]. Both techniques are based on maximum a posteriori (MAP) estimation of Gaussian random variables.

CLUSTER-BASED RECONSTRUCTION

In cluster-based reconstruction of damaged regions of spectrograms, each spectral vector in the spectrogram is assumed to be independent of every other vector. The distribution of the spectral vectors of clean speech is assumed to be a Gaussian mixture, given by

$$P(X) = \sum_{\nu} c_\nu (2\pi|\Theta_\nu|)^{-d/2} e^{-0.5(X-\mu_\nu)^\top \Theta_\nu^{-1}(X-\mu_\nu)}, \tag{11}$$

where d is the dimensionality of the vector, and c_ν, μ_ν, and Θ_ν are, respectively, the a priori probability, mean vector, and the covariance matrix of the νth Gaussian. The Θ_ν matrices are assumed to be diagonal. The distribution parameters are learned from a training corpus of clean speech using the expectation maximization (EM) algorithm [34].

Let Y be any observed spectral vector with some unreliable components, and let X be the corresponding idealized clean vector. Let Y_r, Y_u, X_r and X_u be vectors formed from the reliable and unreliable components of Y and X, respectively. X_r is identical to Y_r; X_u is unknown, but known to be less than Y_u. Ideally, we would estimate the value X_u with the bounded MAP estimator

$$\hat{X}_u = \arg\max_{X_u} \{P(X_u|X_r, X_u \leq Y_u)\}. \tag{12}$$

The probability density of $X = X_r \cup X_u$ is the Gaussian mixture given by (5), and MAP estimation of variables with Gaussian mixture densities is difficult. Hence, we approximate the bounded MAP estimate from the Gaussian mixture as a linear combination of Gaussian-conditional bounded MAP estimates:

$$\hat{X}_u = \sum_{\nu} P(\nu|X_r, X_u \leq Y_u) BMAP(X_u|X_r, X_u \leq Y_u; \mu_\nu, \Theta_\nu). \tag{13}$$

$P(\nu|X_r, X_u \leq Y_u)$ is given by

$$P(\nu|X_r, X_u \leq Y_u) = \frac{c_\nu P(X_r, X_u \leq Y_u|\nu)}{\sum_j c_j P(X_r, X_u \leq Y_u|j)}, \tag{14}$$

where $P(X_r, X_u \leq Y_u|j) = P(X_r, -\infty \leq X_u \leq Y_u; \mu_j, \Theta_j)$ and is computed by (7).

COVARIANCE-BASED RECONSTRUCTION

In covariance-based reconstruction, the log-spectral vectors of a speech utterance are assumed to be samples of a stationary Gaussian random process. The a priori information about the clean speech signal is represented by the statistical parameters of this random process, specifically the expected value of the vectors and the covariances between their components. We denote the mean of the kth element of the mth spectral vector $X(m, k)$ by $\mu(k)$ (since the mean is not a function of time for a stationary process), and the covariance between the k_1th element of the mth spectral vector $X(m, k_1)$ and the k_2th element of the $(m + \xi)$th spectral vector $X(m + \xi, k_2)$ by $c(\xi, k_1, k_2)$, and the corresponding normalized covariance by $r(\xi, k_1, k_2)$:

$$\mu(k) = E[X(m, k)]$$
$$c(\xi, k_1, k_2) = E[(X(m, k_1) - \mu_{k_1})(X(m + \xi, k_2) - \mu_{k_2})]$$
$$r(\xi, k_1, k_2) = \frac{c(\xi, k_1, k_2)}{\sqrt{c(\xi, k_1, k_1)(c(\xi, k_2, k_2))}} \tag{15}$$

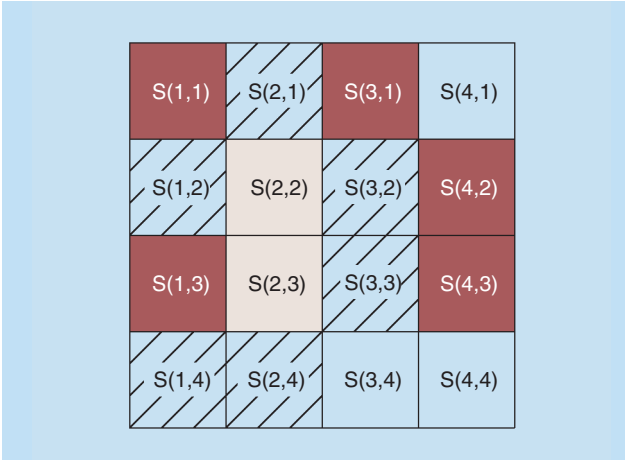
where $E[\cdot]$ represents the expectation operator. These parameters are learned from the spectral vector sequences representing a training corpus of clean speech.

Since the process is assumed to be Gaussian, all unreliable spectral components in any noisy signal can, in principle, be jointly estimated using the bounded MAP procedure. In practice, however, it is equally effective and far more efficient to estimate jointly only the unreliable components in individual spectral vectors, based on the reliable components within a neighborhood enclosing the vectors.

To estimate the unreliable components in the mth spectral vector $X(m)$ we arrange them in a vector $X_u(m)$. We identify the set of all reliable spectral components in the spectrogram that have a normalized covariance of at least 0.5 with at least one of the elements in $X_u(m)$ and arrange them into a neighborhood vector $X_r^{(n)}(m)$. Figure 2 illustrates the construction of the neighborhood vector. The joint distribution of $X_u(m)$ and $X_r^{(n)}(m)$ is Gaussian. The expected values and the covariance matrices of $X_u(m)$ and $X_r^{(n)}(m)$ and the cross covariance between them can all be constructed from the mean and covariance parameters of the underlying Gaussian process. Once the parameters of the joint Gaussian distribution of $X_u(m)$ and $X_r^{(n)}(m)$ are constructed in this fashion, the true value of $X_u(m)$ is estimated using the bounded MAP estimation procedure.

CLASSIFIER-MODIFICATION METHODS

An alternative to estimating the true values of the unreliable or missing components is to modify the classifier itself to perform recognition with unreliable or incomplete data. The two most popular classifier-modification methods have been class-conditional imputation [19] and marginalization [11]. With feature-imputation methods, recognition can be performed using feature vectors that may be different from (and derived from) the spectral vectors that are reconstructed from the partially degraded input. In classifier-modification methods, however, recognition is usually performed using the spectral vectors themselves as features.



[FIG2] Illustration of a hypothetical spectrogram containing four spectral vectors of four components each. Components shown in gray are considered to be unreliable. The blocks with the diagonal lines represent components with a normalized covariance of 0.5 or greater with either S(2, 2) or S(2, 3). To reconstruct unreliable components of the second vector using covariance-based reconstruction, $X_u(2)$ is constructed from S(2, 2) and S(2, 3), and $X_r^{(n)}(2)$ is constructed from S(1, 2), S(1, 4), S(2, 1), S(2, 4), S(3, 2), and S(3, 3).

CLASS-CONDITIONAL IMPUTATION

In class-conditional imputation, state-specific estimates are derived for unknown spectral components, prior to estimating the state output density $P(X|s)$. In other words, when computing the state output density value at any state s for any vector X, the unreliable components of X are estimated using the state output density of s.

We assume that the output density of any state s is a Gaussian mixture, given by

$$P(X|s) = \sum_{\nu} c_{s,\nu} (2\pi |\Theta_{s,\nu}|)^{-\frac{d}{2}} e^{-\frac{1}{2}(X-\mu_{s,\nu})^{\top}\Theta_{s,\nu}^{-1}(X-\mu_{s,\nu})}, \quad (16)$$

where the subscript s applied to the parameters of the distribution indicates that they are specific to the output distribution of state s. As before, let Y be any observed spectral vector, with reliable and unreliable subvectors Y_r and Y_u. Let X represent the corresponding idealized clean vector and X_r and X_u its subvectors corresponding to Y_r and Y_u. The state output density of any state s is defined over the complete vector $X = X_r \cup X_u$. In order to compute $P(X|s)$ we obtain the state-specific estimate for X_u, $\hat{X}_{s,u}$ as

$$\hat{X}_{s,u} = \sum_{\nu} P(\nu|X_r, X_u \le Y_u) \times BMAP(X_u|X_r, X_u \le Y_u; \mu_{s,\nu}, \Theta_{s,\nu}), \quad (17)$$

using the cluster-based reconstruction technique in conjunction with the state output distribution of s. The complete vector $\hat{X}_s = X_r \cup \hat{X}_{s,u}$ is then constructed (with the components arranged in appropriate order), and the state output density value of s is computed as $P(\hat{X}_s|s)$.

MARGINALIZATION

The principle behind marginalization is to perform optimal classification (i.e., recognition) directly based on the observed values of the reliable and unreliable input data. As applied to HMM-based speech recognition systems, this implies that state output densities are now replaced by a term that computes

$$\hat{P}(X|s) = P(X_r, -\infty \le X_u \le Y_u|s). \quad (18)$$

We assume that the state output density $P(X|s)$ for state s is given by (16), and we further assume that all the Gaussians in the mixture density in (16) have diagonal covariance matrices. For any observed spectral vector Y, let R represent the set of

indices of the reliable components and U represent the indices of the unreliable components. In other words, Y_r is composed of all components $Y(k)$ such that $k \in R$, and Y_u is composed of all $Y(k)$ such that $k \in U$. It is easy to show that $\hat{P}(X|s)$ as defined by (18) is now given by

$$\hat{P}(X|s) = \sum_{\nu} c_{s,\nu} P(X_r, -\infty \leq X_u \leq Y_u; \mu_{s,\nu}, \Theta_{s,\nu}), \quad (19)$$

where $P(X_r, -\infty \leq X_u \leq Y_u; \mu_{s,\nu}, \Theta_{s,\nu})$ is given by (7). Note that if the damaged spectral components X_u were assumed to be completely unknown (so that Y_u did not provide any information that constrained X_u), then the limits on the integral terms on the right hand side of (19) would be $(-\infty, \infty)$, and the integral terms would evaluate to 1.0, thereby marginalizing out the unreliable spectral components from the state output density entirely. On the other hand, if it were possible to provide tighter bounds on the true value of X_u, the integral term could use them as limits instead of $(-\infty, Y_u)$.

IDENTIFICATION OF UNRELIABLE COMPONENTS

The most difficult aspect of missing feature methods is the estimation of the spectrographic masks that identify unreliable spectral components. The estimation can be performed in multiple ways: we may either attempt to estimate the SNR of each spectral component to identify unreliable components, or we may attempt to classify unreliable components directly using some other criteria in place of SNR. In the latter case, spectrographic masks may either be estimated from Bayesian principles applied to a variety of measurements derived from the speech signal, or from perceptually-motivated criteria. We discuss all of these alternatives below.

ESTIMATING SPECTROGRAPHIC MASKS BASED ON SNR

To estimate the SNR of spectral components, an estimate of the power spectrum of the corrupting noise is required. Typically, the noise power spectrum is estimated from regions of the signal that are determined to not contain speech. For example, in Vizinho et al. [36], the first several frames of any utterance are assumed to be regions of silence and the average power spectrum of these frames is assumed to represent the power spectrum of the noise. Alternately, the noise power spectrum can be estimated continuously by a simple recursion, in order to track slowly-varying noise. The noise power spectrum is initialized to the average power spectrum of the initial few frames of the incoming signal. Any subsequent sudden increases in energy in the noisy speech signal are assumed to indicate the onset of speech, while regions in the speech whose energies fall below a given threshold are assumed to consist only of noise. Let $Y_p(m, k)$ and $N_p(m, k)$ represent the kth frequency band of the *power* spectra of the observed noisy speech and the noise respectively, in the mth analysis window. The estimated value of $N_p(m, k)$ is obtained as

$$\hat{N}_p(m,k) = \begin{cases} (1-\lambda)\hat{N}_p(m-1,k) + \lambda Y_p(m,k), \\ \quad \text{if } Y_p(m,k) < \beta \hat{N}_p(m,k) \\ \hat{N}_p(m-1,k), \\ \quad \text{otherwise.} \end{cases} \quad (20)$$

Typical values of λ and β are 0.95 and 2, respectively, and the value of λ can be manipulated to track slower or faster variations in the noise. Other noise estimation techniques [18] may also be used in lieu of (20).

Multiple criteria have been proposed to identify unreliable spectral components from the estimated noise spectrum. El-Maliki and Drygajlo [15] propose a negative energy criterion in which a spectral component is assumed to be unreliable if the energy in that component is less than the estimated noise energy in it. In other words $Y_p(m, k)$ is assumed to be unreliable if

$$|Y_p(m,k)| \leq |\hat{N}_p(m,k)|. \quad (21)$$

The SNR criterion, on the other hand, identifies spectral bands as unreliable if the estimated SNR of any spectral component lies below 0 dB. To estimate the SNR, an estimate of the power spectrum of clean speech is required. This is obtained by spectral subtraction, i.e., by subtracting the estimated power spectrum of the noise from the power spectrum of the noisy speech signal as follows [7]:

$$\hat{X}_p(m,k) = \begin{cases} Y_p(m,k) - \hat{N}_p(m,k), \\ \quad \text{if } Y_p(m,k) - \hat{N}_p(m,k) > \gamma Y_p(m,k) \\ \gamma Y_p(m,k), \\ \quad \text{otherwise,} \end{cases} \quad (22)$$

where $\hat{X}_p(m, k)$ represents the estimate of the kth spectral component of the power spectrum of the clean speech signal in the mth analysis window. The parameter γ is a small flooring factor meant to ensure that the estimated power spectrum for the clean speech does not go negative. The SNR criterion states that a spectral component $Y_p(m, k)$ is to be assumed unreliable if the estimated power spectrum of the underlying clean speech is lower than the power spectrum of the noise, i.e., if $\hat{X}_p(m, k) < \hat{N}_p(m, k)$. Alternately stated, $Y_p(m, k)$ is deemed unreliable if less than half the energy in $Y_p(m, k)$ is attributable to speech, i.e., if

$$\hat{X}_p(m,k) < 0.5 Y_p(m,k). \quad (23)$$

In practice, the best mask estimates are obtained when both the negative energy criterion of (21) and the SNR criterion of (23) are used to identify unreliable spectral components. Figure 3(b) shows an example of a spectrographic mask obtained by noise estimation. In general, noise-estimate-based estimation of spectrographic masks may be expected to be

effective when the corrupting noises are stationary or pseudo-stationary. For nonstationary and transient noises, however, the estimation of the noise spectrum is difficult and this technique can result in highly inaccurate spectrographic masks.

BAYESIAN ESTIMATION OF SPECTROGRAPHIC MASKS

The Bayesian approach to estimating spectrographic masks treats the tagging of spectral elements as reliable versus unreliable as a binary classification problem. Renevey et al. [31] use estimates of the distribution of noise to compute an explicit probability that the noise energy in any spectral component exceeds the speech energy in it. In this article, however, we describe an alternative technique presented by Seltzer et al. [35] that does not depend entirely on explicit characterization of the noise. Here, a set of features is computed for every time-frequency location of the spectrogram. Features are designed that exploit the characteristics of the speech signal itself, rather than measurements of the corrupting noise. These features are then input to a conventional Bayesian classifier to determine whether a specific time-frequency component is reliable.

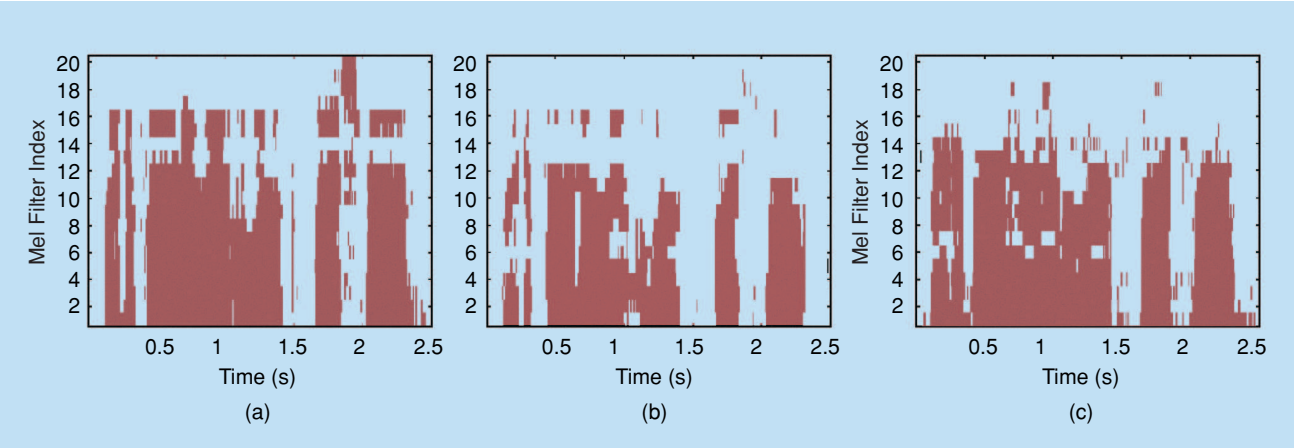
We note that each time-frequency location in the spectrogram actually represents a window of time and a band of frequencies. The features extracted for any time-frequency location are designed to represent the characteristics of the signal components within the corresponding frequency band, in the given window of time. The features for any time-frequency location (m, k) include 1) the ratio of the first and second autocorrelation peaks of the signal within that window, 2) the ratio of the total energy in the kth frequency band to the total energy of all frequency bands, 3) the kurtosis of the signal samples within the mth frame of speech, 4) the variance of the spectrographic components adjoining (m, k), and 5) the ratio of the energy within $Y_p(m, k)$ to the estimated energy of the noise $N_p(m, k)$. The noise estimate is obtained using the procedure outlined in (20). Note that the estimated SNR is only one of the features used and is not the sole determinant of reliability. In addition to the features described above, for voiced regions of speech, an additional feature is derived that measures the ratio of the energy of the signal at all harmonics of the pitch frequency that lie within the kth frequency band [for a location (m, k)], to the energy in frequencies between the pitch harmonics. The pitch estimate required by this feature is obtained with a pitch estimation algorithm.

Separate classifiers are trained for voiced and unvoiced speech, and for each frequency band, using training data for which the true spectrographic masks are known (e.g., clean speech signals that have been digitally corrupted by noise). All distributions are modeled as a mixture of Gaussians, the parameters of which are learned from the feature vectors of appropriately labeled spectral components of the training data (e.g., the distribution of reliable voiced components in the kth frequency band is learned from the feature vectors of reliable components in the kth frequency band of voiced portions of the training data). The a priori probabilities of the reliable and unreliable classes for each classifier are also learned from training data.

During recognition, noisy speech recordings are first segregated into voiced and unvoiced segments (typically using a pitch detection algorithm; segments where it fails to estimate a reasonable pitch value are assumed to be unvoiced). The appropriate feature vectors are then computed for each time-frequency location. The (m, k)th time-frequency location is classified as reliable if

$$P_{V,k}(\text{reliable})P_{V,k}(F(m,k)|\text{reliable}) > P_{V,k}(\text{unreliable})P_{V,k}(F(m,k)|\text{unreliable}) \quad (24)$$

where $F(m, k)$ is the feature vector for (m, k), V is a voicing tag that indicates whether $F(m, k)$ belongs to a voiced segment or not, $P_{V,k}(\text{reliable})$ represents the a priori probability that the kth frequency component of a spectral vector with voicing tag V is reliable, and $P_{V,k}(F|\text{reliable})$ represents the distribution of feature vectors of reliable components in the kth frequency band of speech segments with voicing tag V.



[FIG3] (a) An ideal spectrographic mask for an utterance corrupted to an SNR of 10 dB by white noise. Reliable time-frequency components have been identified based on their known SNR. An SNR threshold of 0 dB has been used to generate this mask. (b) Spectrographic mask for the same utterance obtained by estimating the local SNR of the signal. (c) Spectrographic mask obtained by using Bayesian classification.

Figure 3(c) shows an example of a spectrographic mask obtained using the Bayesian approach. We observe that the Bayesian mask is superior to the noise-estimate-based mask in this example. In general, since Bayesian mask estimation is not strictly dependent on the availability of good estimates of the noise spectrum (as the estimated SNR is only one of the features used), it is effective both for stationary and nonstationary noises. For example, Seltzer et al. [35] report that effective spectrographic masks can be derived using the Bayesian approach for speech corrupted by music, whereas masks derived from noise estimates are totally ineffective. Nevertheless, Bayesian mask estimation may be expected to fail when the spectral characteristics of the noise are similar to those of speech, such as when the corrupting signal is speech from a competing speaker. The Bayesian mask also has the advantage that the classifier computes the probability of unreliability $P(\text{unreliable}|F(m, k))$. These probabilities can be used in conjunction with the soft-mask technique described later in this article.

THE ACCURACY OF AUTOMATIC SPEECH RECOGNITION SYSTEMS IS STILL ADVERSELY AFFECTED BY NOISE AND OTHER SOURCES OF ACOUSTICAL VARIABILITY.

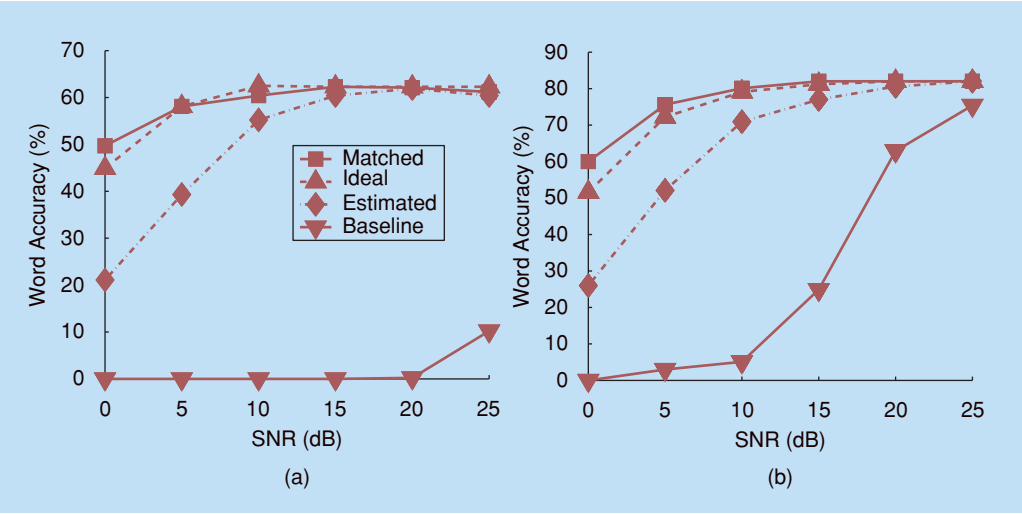
MASK ESTIMATION FROM PERCEPTUAL CRITERIA

This approach attempts to identify groups of speech-like spectral components based on the physics of sound and selected properties of the human auditory system. For example, in voiced speech, most of the energy tends to occur around the harmonics of the fundamental frequency. Barker et al. [4] propose that within each frame of speech that is identified as voiced and has a valid pitch estimate, all time-frequency components that occur at the harmonics of the pitch may be assumed to be reliable and represent speech. Such masks, however, have been found to be most effective when used in conjunction with other masks, such as noise-estimate based masks; any spectral component that is identified as reliable by either mask is assumed to be reliable. Palomaki et al. [24] have described a binaural processing model that extracts information about interaural time delays and intensity differences to identify the reliable time-frequency regions in a representation of a complex auditory scene that are believed to arise from a common azimuth. By passing these reliable components to a missing feature recognizer similar to the fragment decoder described below, Palomaki et al. have demonstrated that the use of cues extracted using binaural processing can lead to a very substantial improvement in speech recognition accuracy when the target speech signal and interfering signals arrive from different azimuths.

Spectrographic masks that are derived from perceptual principles are based entirely on the known structural patterns of speech spectra and are not dependent on noise estimates in any manner. As a result, this approach can result in effective spectrographic masks even in very difficult environments, e.g., in the presence of competing speech signals.

DEALING WITH UNCERTAINTY IN MASK ESTIMATION

Regardless of the actual technique employed, it is never possible to identify unreliable spectral components with complete certainty. Errors in the spectrographic mask can cause the recognition performance of missing feature methods to degrade significantly. Figure 4 describes the recognition accuracy obtained with various missing feature methods, comparing the accuracy obtained using oracle spectrographic masks (that are obtained with knowledge of the true SNR of the spectral components) to the accuracy obtained using masks that were estimated blindly from the incoming data. The degradation in recognition



[FIG4] Recognition accuracy for speech corrupted by white noise as a function of SNR using two different missing feature methods. (a) Recognition accuracy obtained using marginalization. Recognition is performed using log-spectral vectors. (b) Recognition accuracy obtained using cluster-based reconstruction. Recognition is performed with cepstral vectors derived from reconstructed spectrograms. In both panels, the triangular symbols show recognition performance obtained with ideal spectrographic masks, while the diamonds show recognition accuracy with estimated spectral masks. The square symbols show the recognition accuracy of a matched recognizer that had been trained in the testing environment, while the delta symbols at the bottom show the accuracy obtained when the recognizer is trained with clean speech.

accuracy due to errors in mask estimation is evident. We have observed that erroneous tagging of reliable components as being unreliable degrades recognition accuracy more than tagging unreliable elements as reliable. More detailed analyzes of the effects of mask estimation errors may be found in [28] and [35].

Two different remedies have been proposed to cope with the effects of mask estimation errors. In the first approach, soft masks are estimated that represent the probability that a given spectral component is reliable (as opposed to a binary mask that unambiguously tags spectral components as reliable or unreliable). In the second approach, the mask estimation is performed as a part of the speech recognition process itself, with the expectation that the detailed and precise statistical models for speech that are used by the recognizer will also enable more accurate estimation of masks.

SOFT MASKS

Under the soft-decision framework, each spectral element $Y(m, k)$ is assigned a probability $\gamma = P(\text{reliable}|Y(m, k))$ that it is reliable and dominated by speech rather than noise. The probability of reliability is arrived at differently by different authors. The Bayesian mask estimation algorithms of Seltzer et al. [35] and Renevey et al. [31] can be used to obtain γ directly. Barker et al. [6] employ an alternate approach, approximating γ as a sigmoidal function of the estimated SNR of $Y(m, k)$.

$$\gamma \approx \frac{1}{1 + \exp(-\alpha(\text{SNR}(m, k) - \beta))}. \tag{25}$$

The SNR itself is estimated from the estimated differences in level between speech and noise, using the procedures described previously. Typical values for α and β are 3.0 and 0.0, respectively.

Morris et al. [22] show that marginalization based on soft masks can be implemented with a relatively minor modification to the bounded marginalization algorithm of (19) for HMM-based speech recognition systems that model state output densities as mixtures of Gaussians with diagonal covariance matrices:

$$P(Y|s) = \sum_k \prod_j \times \left(\gamma \frac{e^{\frac{-(Y(j)-\mu_{s,k}(j))^2}{2\sigma_{s,k}(j)}}}{\sqrt{2\pi\sigma_{s,k}(j)}} + \frac{(1-\gamma)}{Y(l)} \int_0^{Y(l)} \frac{e^{\frac{-(x-\mu_{s,k}(l))^2}{2\sigma_{s,k}(l)}}}{\sqrt{2\pi\sigma_{s,k}(l)}} dx \right). \tag{26}$$

Note that in (26) $Y(l)$ is the lth component of Y; (26) is equivalent to a model that assumes that spectral features are corrupted by a noise process that leaves them unchanged with a probability γ, or with a probability $1 - \gamma$ adds to them a random value drawn from a uniform probability distribution in the range $(0, Y(l))$. We also note that marginalization is the only missing feature technique that can utilize soft masks effectively, since other techniques require unambiguous identification of unreliable spectral components.

SIMULTANEOUS RECOGNITION AND MASK ESTIMATION

Possibly the most sophisticated missing feature technique is the speech fragment decoder proposed by Barker et al. [3], [5]. In the speech fragment decoder the search mechanism used by the speech recognizer is modified to derive the optimal spectrographic mask jointly with the optimal word sequence. In order to do this, the Bayesian classification equation given in (5) is modified to become

$$\hat{W}, \hat{\mathbf{S}} = \arg\max_{W,\mathbf{S}} \{P(W, \mathbf{S}|\mathbf{Y})\} \tag{27}$$

where $\mathbf{S}$ represents any spectrographic mask and $\hat{\mathbf{S}}$ represents the optimal mask, and $\mathbf{Y}$ represents the entire spectral vector sequence for the noisy speech. Letting $\mathbf{X}$ be the ideal spectral vector sequence for the clean speech underlying Y, this can be shown to be equal to

$$\hat{W}, \hat{\mathbf{S}} = \arg\max_{W,\mathbf{S}} \left\{ P(W) \left(\int P(\mathbf{X}|W) \frac{P(\mathbf{X}|\mathbf{S}, \mathbf{Y})}{P(\mathbf{X})} dX \right) P(\mathbf{S}|\mathbf{Y}) \right\} \tag{28}$$

where $P(\mathbf{S}|\mathbf{Y})$ represents the probability of spectrographic mask $\mathbf{S}$ given the noisy spectral vector sequence $\mathbf{Y}$, which is assumed to be independent of the word sequence W. In practical implementations of HMM-based speech recognition systems, the optimal state sequence is determined along with the best word sequence. The corresponding modification to (27) used by the speech fragment decoder is

$$\hat{W}, \hat{\mathbf{S}} = \arg\max_{W,\mathbf{S}} \max_s \left\{ P(\mathbf{S}|\mathbf{Y}) P(W) P(s|W) \times \left(\prod_m \int P(X(t)|s_m) \frac{P(X(t)|\mathbf{S}, \mathbf{Y})}{P(X(m))} dX(m) \right) \right\}. \tag{29}$$

Here s represents a state sequence through the HMM for W, s_m represents the mth state in the sequence, $P(s|W)$ represents the a priori probability of s and is obtained from the transition probabilities of the HMM for W, $X(m)$ represents the mth vector in $\mathbf{X}$, and $P(X|s)$ represents the state output density of s.

Let $S(m)$ represent the spectrographic mask vector for $Y(m)$. Let $U(m)$ represent the set of spectral components of $Y(m)$ that are tagged as unreliable (or, alternately, as belonging to the background noise) by $S(m)$. Let $R(m)$ be the set of frequency bands that are identified as reliable. Let $Y(m, j)$ and $X(m, j)$ represent the jth component of $Y(m)$ and $X(m)$,

respectively. All state output densities are modeled as mixtures of Gaussians with diagonal covariance matrices, as given by (16). By further assuming that the conditional probabilities $P(X(m, j)|\mathbf{S}, \mathbf{Y})$ are independent for the various values of j, and that $P(X(m, j)|\mathbf{S}, \mathbf{Y})$ is 0 for $X(m, j) > Y(m, j)$ and proportional to $P(X(m, j))$ otherwise, and finally that the a priori probability of $X(m, j)$ is a uniform density between 0 and $x_{\max}$, the acoustic probability term within the parentheses on the right hand side of (29) is computed as

$$\int P(X(m)|s_m)\frac{P(X(m)|\mathbf{S}, \mathbf{Y})}{P(X(m))}dX(m) = \sum_k c_{s_t,k} \prod_{j\in R(m)} e^{\frac{-(Y(m,j)-\mu_{s,k}(j))^2}{2\sigma_{s,k}(j)}} \prod_{l\in U(m)} \int_0^{Y(m,l)} e^{\frac{-(x-\mu_{s,k}(j))^2}{2\sigma_{s,k}(j)}} \frac{x_{\max}}{Y(m, l)}dx \qquad (30)$$

A complete implementation of (29) requires exhaustive evaluation of every possible spectrographic mask and is computationally infeasible. Instead, the fragment decoder hypothesizes a large number of fragments or regions of the spectrogram within which all components are assumed to belong together. Fragments may be hypothesized based on various criteria including SNR and acoustic cues such as harmonicity. Joint recognition and search for the optimal spectrographic mask is performed only over spectrographic masks that may be formed as combinations of these fragments. Barker et al. implement this procedure using a token-passing algorithm that is described in [3].

The fragment-decoding approach has been shown by Barker et al. to be more effective than simple bounded marginalization. It also has the advantage over other missing feature methods that it can be easily extended to recognize multiple mixed signals, such as those obtained from the combined speech of multiple simultaneous talkers. The only requirement is that the precomputed fragments group spectral components from a single speaker accurately.

ADDITIONAL PROCESSING OF SPECTRAL INPUT

The discussion thus far has assumed implicitly that recognition is performed directly using the incoming spectral vectors as features. However, most speech recognizers preprocess incoming feature vectors in various ways in order to improve recognition accuracy. For example, recognizers usually use cepstra derived from log spectral vectors, rather than the log-spectral vectors themselves, because cepstral vectors are known to result in significantly greater accuracy [13]. Recognition systems that use feature-vector imputation can work from cepstral vectors since these can be derived from the complete spectral vectors reconstructed by this approach. On the other hand, classifier-modification methods are generally ineffective for recognizers that work from cepstral vectors, since they require information that characterizes the reliability of each component of the incoming feature vectors, and such information is available only for spectral vectors.

Other common types of preprocessing of incoming feature vectors include augmentation using difference vectors and mean normalization of feature vectors. Again, while these manipulations do not pose problems for feature-vector reconstruction methods, classifier-modification methods must be modified to accommodate them, and they in turn constrain the specific missing feature methods that can be employed.

TEMPORAL-DIFFERENCE FEATURES

In most recognizers, the basic feature vector at any time is augmented by difference and double-difference vectors. Difference vectors represent the trend, or velocity, of the feature vectors and are usually obtained as the difference between adjacent feature vectors. The use of this information partially compensates for the commonly used but clearly inaccurate assumption that features extracted from successive analysis frames are statistically independent. The difference vector for frame m is obtained as

$$D(m) = Y(m + \xi) - Y(m - \xi) \qquad (31)$$

where $D(m, k)$, the kth component of $D(m)$ becomes unreliable if either $Y(m + \xi, k)$ or $Y(m - \xi, k)$ is unreliable. Double difference vectors represent the acceleration of the spectral vectors and are computed as the difference between adjacent difference vectors:

$$DD(m) = D(m + \zeta) - D(m - \zeta) \qquad (32)$$

$DD(m, k)$ is unreliable if either $D(m + \zeta, k)$ or $D(m - \zeta, k)$ is unreliable. In the worst case, difference vectors may have as much as twice as many unreliable components on average as the spectral vectors, while double difference vectors may have up to four times as many unreliable components. In practice, however, unreliable components tend to occur in contiguous patches, and the fraction of components in the difference and double difference vectors that are unreliable tends not to be much greater than those of the spectral vectors themselves.

The upper and lower bounds on the values of unreliable difference and double difference vector components must be computed from the bounds on (or values of) the spectral components of which they are composed. For methods that use soft masks, the reliability probability of these terms must also be derived from the reliability probabilities of the spectral components from which they were computed.

MEAN NORMALIZATION

Mean normalization [2] refers to the procedure by which the mean feature vector of any sequence of feature vectors is subtracted from all vectors in the sequence, so that the resulting

sequence has zero mean. In symbolic terms, the mth normalized vector $\bar{Y}(m)$ of any sequence of feature vectors $Y(1), Y(2), \ldots, Y(M)$ is computed as

$$\bar{Y}(m) = Y(m) - \frac{1}{M}\sum_{m} Y(m). \tag{33}$$

Mean normalization is commonly observed to result in large improvements in the recognition accuracy of automatic speech recognitions systems. Unfortunately, mean normalization as described by (33) cannot be performed with classifier-modifying missing feature methods because many components of the feature vectors used for recognition are potentially unreliable. The mean value of all vectors, as used in (33), would include the contributions of these unreliable components and would be unreliable itself, and normalization by such a mean estimate results in degradation of recognition performance [30]. Alternately, the mean value could be computed from only the reliable components of the spectrogram. Unfortunately, the reliable regions of spectrograms contain chiefly high-energy spectral components, and mean estimates obtained from them tend to be biased, again resulting in degraded recognition performance.

MISSING FEATURE METHODS MODEL THE EFFECTS OF NOISE ON SPEECH AS THE CORRUPTION OF REGIONS OF TIME-FREQUENCY REPRESENTATIONS OF THE SPEECH SIGNAL.

A useful substitute for mean normalization has been proposed by Palomaki et al. [25]. Instead of subtracting the *mean* value of the spectral vector, every spectral component is normalized by a value that represents the upper percentile of the values for that component. Using this approach, $\bar{Y}(m,k)$, the normalized value of $Y(m,k)$, is computed as

$$\bar{Y}(m,k) = Y(m,k) - \frac{1}{D}\sum_{\tau \in L} Y(\tau,k) \tag{34}$$

where L represents the set of frame indices of the D greatest $Y(m,k)$ values that are reliable. Palomaki et al. [25] observe that five is a good value for D, although lower values are also effective. This approach has the advantage that the normalization term does not get biased by the presence of unreliable components in the spectrum. Thus, the normalization given by [34] can be effectively used with missing feature methods.

EXPERIMENTAL RESULTS AND DISCUSSION

In this section we discuss and compare various aspects of missing feature methods and their relative merits. Where possible, we present experimental evidence in support of our statements. We note that missing feature methods remains an active area of research and that the techniques presented in this article have been developed by a number of people from a variety of research groups. Since not all of these researchers have worked on all aspects of the problem, it is not possible to identify a consistent set of results that have all been obtained on the same systems and databases. Consequently, the experimental results we present in this article have been culled from a number of papers and sources. We have attempted to maintain a modicum of consistency where possible; most of the results described have been obtained from experiments conducted in the authors' laboratory at Carnegie Mellon University (CMU). The CMU experiments were conducted using the Resource Management database [26], with the Sphinx-3 HMM-based speech recognition system, with 2,000 tied state distributions with Gaussian state output densities. We have mainly presented results for speech corrupted by white noise, although similar results have also been obtained using other noise types. Other results reported in this article have been drawn from experiments conducted at the University of Sheffield and elsewhere, using the TI digits database. Where we report such results, the details of the experimental setup used for the experiments have been provided.

Speech is a highly redundant signal, with the evidence for any acoustic event being multiply represented in several frequency bands, often over several tens of milliseconds of time. Raj et al. [28] report that speech recognition performance does not degrade significantly when a randomly selected 80% of the elements are excised from a spectrogram and recognition is performed with only the remaining elements. Similar results have been reported by other researchers, e.g., Cooke et al. [9]. The missing feature approach therefore promises to be highly effective for noise-robust speech recognition. Traditionally, the best strategy to recognize speech that has been corrupted by stationary noise has been to train a matched recognizer with speech that had been corrupted to the same level by the same kind of noise. Figure 4(a) and (b) compares the recognition accuracy obtained using matched recognizers on speech corrupted to various SNRs by white noise, with the performance obtained with two different missing feature methods using ideal spectrographic masks that had been derived with perfect knowledge of the true SNR of every spectrographic element, as well as using recognizers that had been trained with clean speech. The recognition accuracy performance obtained with the missing feature method is observed to be comparable to that obtained with a matched recognizer.

In practice, spectrographic masks must be estimated, and the recognition accuracy obtained with estimated masks (also shown in Figure 4) is significantly worse than that obtained using ideal masks. Nevertheless, missing feature methods have the advantage that the recognizer need not be retrained for every noise type or level, and they also hold the promise that the performance could improve significantly with improved mask estimation. More importantly, matched recognizers can-

not be trained for most practical situations, since the level and characteristics of the noise change even within the course of an utterance. Instead, multistyle recognizers are trained that attempt to strike an effective compromise across all the observed noise types and levels. Experiments reported by Barker et al. [4] (Figure 5) show that missing feature methods can outperform such recognizers even when the spectrographic masks are estimated.

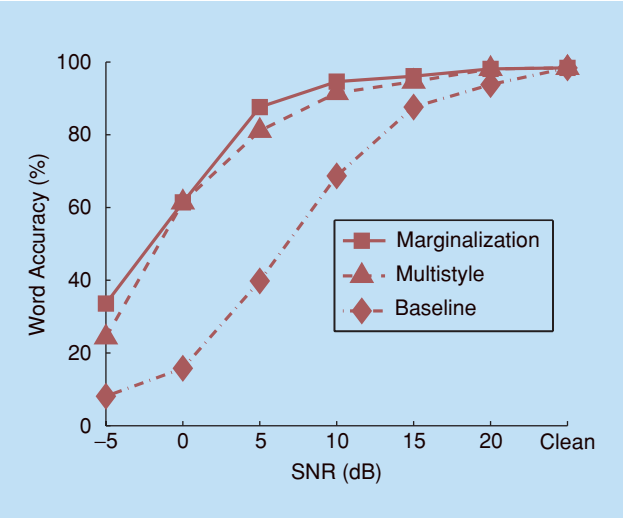
As noted earlier, highly nonstationary noises such as music pose special problems for speech recognition systems. Conventional noise compensation schemes are rendered ineffective by such noises. However, missing feature methods have been shown to result in significant improvements in recognition accuracy even on such noises (Figure 6).

THE MOST DIFFICULT ASPECT OF MISSING FEATURE METHODS IS THE ESTIMATION OF THE SPECTROGRAPHIC MASKS THAT IDENTIFY UNRELIABLE SPECTRAL COMPONENTS.

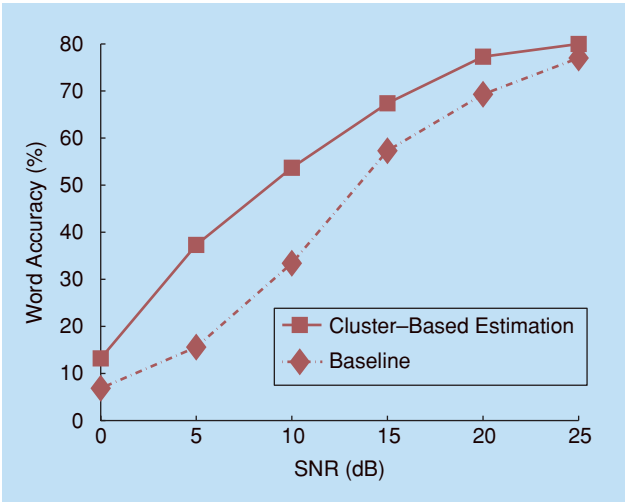
Not all missing feature methods are equivalent and useful in all situations, however, and different implementations have different characteristics. The SNR threshold that is used to determine which signal components are unreliable is typically higher for classifier-modification methods than for feature-imputation methods. For missing feature methods with relatively low SNR thresholds for unreliability, even spectral components tagged as reliable often have a certain degree of noise. For such methods, reducing the noise level in these components using a technique such a spectral subtraction improves recognition accuracy [28]. For methods for which the SNR threshold for reliability is relatively high, spectral subtraction does not greatly improve accuracy.

Of the various missing feature methods, only marginalization and its derivative algorithms (soft-mask marginalization and fragment decoding) purport to perform optimal classification. Consequently the best recognition accuracy can be expected from marginalization. Unfortunately, marginalization has the disadvantage that recognition must be performed with spectral vector sequences directly. In general, however, recognition accuracy obtained with cepstral vectors derived from spectral vectors is much superior to that obtained with spectral vectors. Feature-imputation methods that reconstruct entire spectrograms enable recognition with cepstral vectors derived from the reconstructed spectrograms. The benefits of going to the cepstral domain often overcome the advantage gained by the optimal classification performed with spectral vectors in marginalization, especially at high SNRs. This is demonstrated in Figure 7, which compares recognition performance obtained with spectral and cepstral vectors using various missing feature methods.

Figure 7 and other results reported in the literature also shows that class-conditional imputation and covariance-based reconstruction generally result in much poorer recognition than either marginalization or cluster-based reconstruction. However, these methods do have their uses. For example, Josifovski et al. [19] report that class-conditional imputation is highly effective for reconstruction of complete spectrograms from partial or unreliable ones. Raj et al. [28] show that when spectrographic components are lost due to random deletions (e.g., due to loss during transmission), covariance-based reconstruction, which draws on information from adjacent vectors to reconstruct any missing component, is by far the best spectrogram reconstruction method.

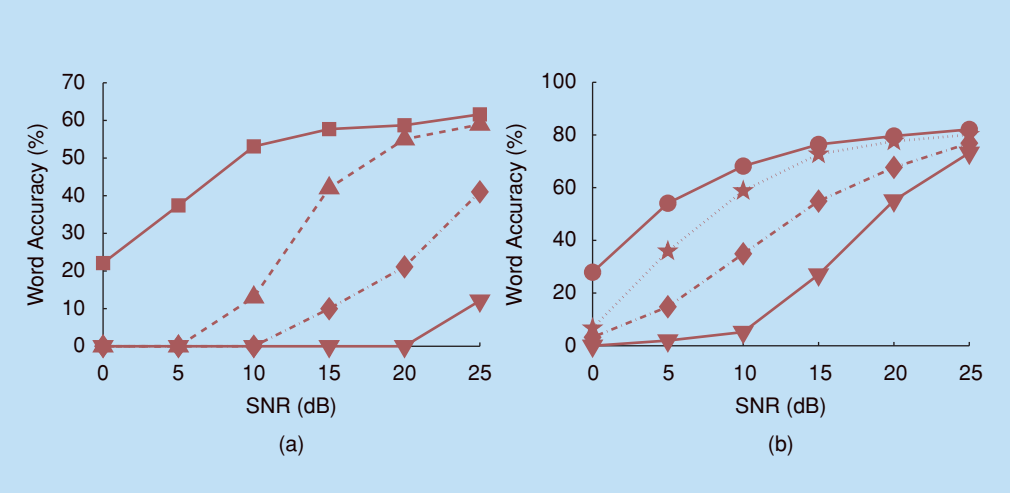


[FIG5] Comparison of recognition accuracy obtained using multistyle training and with soft-mask marginalization. The experiment was conducted on Test Set A of the Aurora corpus. Recognition was performed with spectral vectors.



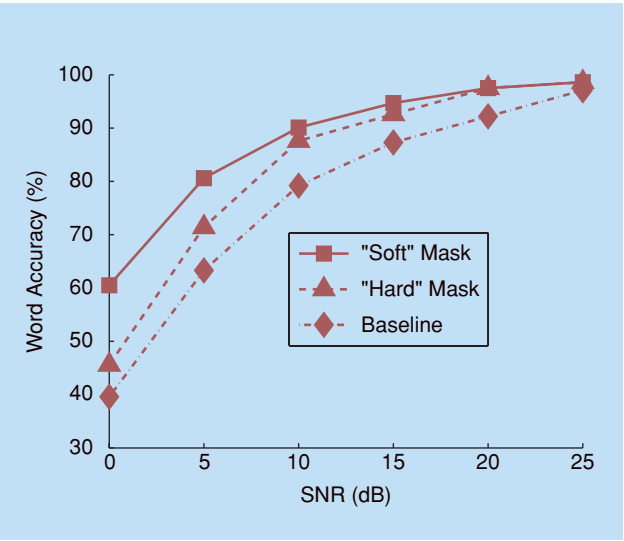
[FIG6] Recognition accuracy for speech corrupted by music. The spectrographic mask was estimated using a Bayesian classifier, and unreliable components were reconstructed by cluster-based reconstruction. Recognition was performed using cepstra derived from reconstructed spectrograms. The lower curve shows the recognition accuracy obtained when no missing feature methods were used. The RM database was used for this experiment, and the recognizer was trained with clean speech.

In considering the data in Figure 7 it should be noted that the relatively poor recognition results obtained using spectral coefficients may be partially accounted for by the fact that recognition was performed using 20-dimensional log-spectral features, with single-Gaussian HMM output distributions and no cepstral mean normalization or similar processing. Cepstral mean normalization was not employed in these experiments because it cannot be meaningfully applied for the marginalization method for the reasons discussed earlier. Cooke and his colleagues have suggested that the difference between the recognition accuracy obtained with spectral and cepstral vectors may be greatly reduced (although perhaps not eliminated) by using more detailed state output distributions for the HMMs. They have also noted that cepstral vectors are intrinsically able to provide a greater degree of normalization for level and spectral tilt than log-spectral features computed without mean normalization.

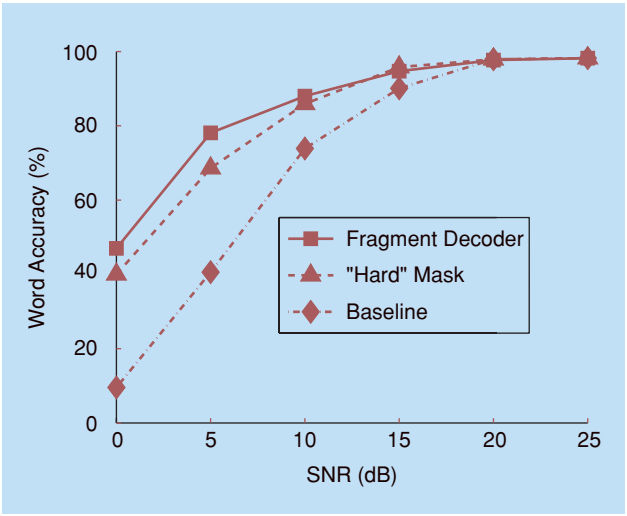


[FIG7] Recognition accuracy obtained with various missing feature methods on speech corrupted by white noise. (a) shows the performance obtained with classifier-modification methods on a recognizer that works with spectral vectors. Recognition accuracy is obtained using marginalization (square symbols) and state-based imputation (triangles). (b) shows recognition performance obtained with cepstra derived from spectrograms reconstructed by the feature-imputation methods of cluster-based imputation (circles) and correlation-based imputation (stars). These results are compared with results obtained using front ends that do not perform missing feature analysis: spectral subtraction (diamonds), and baseline performance (deltas).

A serious problem for missing feature methods is that mask estimation is an unreliable process, and estimated masks are often errorful. Of all the missing feature methods, marginalization is the most robust to mask estimation errors. The soft-mask variant of marginalization described previously can result in greater robustness to uncertainty in mask estimation, as demonstrated by the results in Figure 8 obtained by Barker et al. [6]. The fragment decoder described previously is, in principle, the



[FIG8] Comparison of recognition accuracy obtained using soft-mask marginalization (squares) with the accuracy obtained using conventional bounded marginalization (triangles) on speech corrupted by Lynx helicopter noise. The TI digits corpus has been used for this experiment. Recognition was performed with spectral vectors. The baseline performance obtained without missing-feature methods is also shown.



[FIG9] Comparison of recognition accuracy obtained by fragment decoding (squares) with that obtained by bounded marginalization (triangles). Recognition was performed with spectral vectors for both missing feature methods. Baseline recognition accuracy with mel frequency cepstral coefficients, using cepstral mean normalization, is also shown. The TI-digits database was corrupted by noise samples from the NOISEX database for this experiment.

optimal method for mask estimation since it actually uses the recognizer itself to identify the best mask and can result in significant improvements over the recognition accuracy obtained when spectrographic masks are computed separately from the recognizer. Typical results are illustrated in Figure 9, using data provided by Barker et al. [3]. In practice, its performance is limited by the accuracy of the hypothesized fragments that are themselves obtained using other, potentially errorful techniques. These errors can be reduced by hypothesizing smaller fragments with greater likelihood of consistency, at the cost of increased computation.

The inability to perform additional processing such as mean subtraction has sometimes been considered a problem for classifier-modification-based missing feature methods. Nevertheless, the technique proposed by Palomaki et al. [25] described previously provides a good substitute for CMN that can be used with missing feature methods. Unfortunately, no mechanism has been developed to take advantage of either this technique or difference vectors for soft-mask methods or fragment decoders.

The dramatic and consistent success enjoyed by the missing feature techniques described in this article may cause one to question why these approaches have not been widely employed in state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems such as those developed for the Defense Advanced Research Projects Agency (DARPA) Broadcast News and Switchboard tasks. While there are probably a number of reasons for this (including the observations that some algorithms are somewhat computationally costly and that the field has only stabilized significantly in recent years), it is certainly the case that the speech signals encountered in tasks like Broadcast News and Switchboard are more degraded by the effects of speaking style and disfluencies in speech production than they are by background noise. We expect that missing feature techniques will be employed much more widely in tasks for which there is significant degradation due to noise and especially if the sources in noise are nonstationary or transient in nature.

A final topic that has not been explored significantly is the issue of training with noisy data. Conventional recognizers have benefited greatly when they have been trained with the kind of noisy speech that they are expected to recognize. Missing feature methods, on the other hand, are usually developed within the paradigm of training the recognizer with clean data. It is likely that the performance of these methods would improve further if the recognizer itself has been trained with incomplete or unreliable spectrograms obtained from noisy speech. While the mathematics of such a training procedure are well known, the topic itself remains to be explored.

SUMMARY

In this article we have reviewed a wide variety of techniques based on the identification of missing spectral features that have proved effective in reducing the error rates of automatic speech recognition systems. These approaches have been conspicuously effective in ameliorating the effects of transient maskers such as impulsive noise or background music. We described two broad classes of missing feature algorithms: feature-vector imputation algorithms (which restore unreliable components of incoming feature vectors) and classifier-modification algorithms (which dynamically reconfigure the classifier itself to cope with the effects of unreliable feature components). We reviewed the mathematics of four major missing feature techniques: the feature-imputation techniques of cluster-based reconstruction and covariance-based reconstruction, and the classifier-modification methods of class-conditional imputation and marginalization. We then considered the very difficult task of estimating the spectrographic masks that identify which components of incoming spectral vectors are unreliable, focusing on the estimation of spectrographic masks from estimates of statistical parameters that represent characteristics of the noise, as well as Bayesian estimation methods that typically exploit characteristics of the speech to be recognized. We concluded our discussion of spectrographic masks with a description of how uncertainty in the estimation can be incorporated into the recognition process, describing soft-mask techniques, and fragment-decoding systems that simultaneously recognize the spectrographic masks and the incoming spoken utterance. We also discussed the ways in which the common feature extraction procedures of cepstral analysis, temporal-difference features, and mean subtraction can be handled by speech recognition systems that make use of missing feature techniques. We concluded with a discussion of a small number of selected experimental results. These results confirm the effectiveness of all types of missing feature approaches discussed in ameliorating the effects of both stationary and transient noise, as well as the particular effectiveness of both soft masks and fragment decoding.

ACKNOWLEDGMENTS

We are extremely grateful to Mike Seltzer for his help and advice in many aspects of this work and to Martin Cooke and Jon Barker for permission to use some of their experimental data. We also thank Martin Cooke and Dan Ellis for many helpful comments and suggestions in the review process. The work described was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant N66001-99-1-8905. Preparation of the manuscript is partially supported by the Mitsubishi Electric Research Laboratories and by the National Science Foundation (Award IIS 0420866).

AUTHORS

Bhiksha Raj received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, Pennsylvania, in May 2000. Since 2001, he has been working at Mitsubishi Electric Research Laboratories, Cambridge, Massachusetts. He works mainly on algorithmic aspects of speech recognition, with special emphasis on improving the robustness of speech recognition systems to environmental noise. His latest work was on the use of statistical information

encoded by speech recognition systems for various signal processing tasks. He is a Member of the IEEE.

Richard M. Stern received the S.B. degree from the Massachusetts Institute of Technology (1970), the M.S. degree from the University of California, Berkeley (1972), and the Ph.D. from MIT (1977), all in electrical engineering. He has been on the faculty of Carnegie Mellon University since 1977, where he is currently a professor in the Electrical and Computer Engineering, Computer Science, and Biomedical Engineering Departments and the Language Technologies Institute. Much of his current research is in spoken language systems. He is particularly concerned with the development of techniques to make automatic speech recognition more robust with respect to changes in environmental and acoustical ambience. He has also developed sentence parsing and speaker adaptation algorithms for earlier CMU speech systems. In addition to his work in speech recognition, he also maintains an active research program in psychoacoustics, where he is best known for theoretical work in binaural perception. He is a Member of the IEEE and the Acoustical Society of America, and he was a recipient of the Allen Newell Award for Research Excellence in 1992.

REFERENCES

[1] S. Ahmed and V. Tresp, "Some solutions to the missing feature problem in vision," in *Advances in Neural Information Processing Systems 5*. San Mateo, CA: Morgan Kauffman, 1993, pp. 393–400.

[2] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[3] J. Barker, M.P. Cooke, and D.P.W. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, no. 1, pp. 5–25, 2005.

[4] J. Barker, M. Cooke, and P. Green, "Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. Eurospeech-2001*. Aalborg, Denmark: 2001, pp. 213–216.

[5] J. Barker, M. Cooke, and D. Ellis, "Combining bottom-up and top-down constraints for robust ASR: The multisource decoder," in *Proc. Workshop Consistent Reliable Acoustic Cues (CRAC)*, Sept. 2001.

[6] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP 2000*, Beijing, China, Sept. 2000, pp. 373–376.

[7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[8] M.P. Cooke and D.P.W.E. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Commun.*, vol. 35, no. 3–4, pp. 141–177, 2001.

[9] M.P. Cooke, P.G. Green, and M.D. Crawford, "Handling missing data in speech recognition," in *Proc. ICSLP-1994*, Yokohama, Japan, 1994, pp. 1555–1558.

[10] M.P. Cooke, A. Morris, and P.D. Green, "Missing data techniques for robust speech recognition," in *Proc. IEEE Conf. Acoustics, Speech, Signal Processing*, Munich, Germany, 1997, pp. 863–866.

[11] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust ASR with unreliable data and minimal assumptions," in *Proc., Robust'99*, Tampere, Finland, 1999, pp. 195–198.

[12] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust Automatic Speech Recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267–285, 2000.

[13] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[14] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[15] M. El-Maliki and A. Drygajlo, "Missing features detection and handling for robust speaker verification," in *Proc. Eurospeech 1999,* Budapest, Hungary, pp. 975–978.

[16] B.J. Frey, L. Deng, A. Acero, and T. Kristjansson, "ALGONQUIN: Iterating LaPlace's method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 901–904.

[17] H. Fletcher, *Speech and Hearing in Communication*. New York: Van Nostrand, 1953.

[18] H.G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. IEEE Conf. Acoustics, Speech Signal Processing*, Detroit, Michigan, 1995, pp. 153–156.

[19] L. Josifovski, M. Cooke, P. Green, and A. Vizihno, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2837–2840.

[20] R.P. Lippmann and B.A. Carlson, "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," in *Proc. Eurospeech*, Rhodes, Greece, 1997, pp. KN37–KN40.

[21] B.C.J. Moore and B.R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, no. 3, pp. 750–753, 1983.

[22] A.C. Morris, J. Barker, and H. Bourlard, "From missing data to maybe useful data: Soft data modelling for noise robust ASR," in *Proc. WISP 2001*, Stratford-upon-Avon, U.K., 2001.

[23] P.J. Moreno, "Speech recognition in noisy environments," Ph.D. dissertation, ECE Dept., Carnegie Mellon Univ., Pittsburgh, PA, May 1996.

[24] K.J. Palomaki, G.J. Brown, and D.L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Commun.*, vol. 43, no. 4, pp. 361–378, Sept. 2004.

[25] K.J. Palomaki, G.J. Brown, and J. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Commun.*, vol. 43, no. 1–2, pp. 123–142, 2004.

[26] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallet, "The DARPA 1000 word resource management database for continuous speech recognition," in *Proc. IEEE Conf. Acoustics, Speech Signal Processing*, New York, 1988, pp. 651–654.

[27] L.R. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[28] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. dissertation, ECE Dept., Carnegie Mellon Univ., Pittsburgh, PA, Apr. 2000.

[29] B. Raj, V. Parikh, and R.M. Stern, "The effects of background music on speech recognition accuracy," in *Proc. IEEE Conf. Acoustics, Speech Signal Processing*, Munich, Germany, 1997, pp. 851–854.

[30] B. Raj, M.L. Seltzer, and R.M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun. J.*, vol. 43, no. 4, pp. 275–296, 2004.

[31] P. Renevey and A. Drygajlo, "Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition," in *Proc. EUROSPEECH*, Budapest, Hungary, 1999, pp. 2627–2630.

[32] P. Renevey and A. Drygajlo, "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proc. Workshop Consistent Reliable Acoustic Cues (CRAC)*, Aalborg, Denmark, 2001.

[33] P. Renevey, "Speech in noisy conditions using missing feature approach," Ph.D. dissertation, Swiss Federal Institute of Technology, Lousanne, Switzerland, 2001.

[34] D.B. Rubin, *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

[35] M.L. Seltzer, B. Raj, and R.M. Stern, "A bayesian framework for spectrographic mask estimation for missing feature speech recognition," *Speech Commun. J.*, vol. 43, no. 4, pp. 379–393, 2004.

[36] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study," in *Proc. Eurospeech*, Budapest, Hungary, 1999, pp. 2407–2410. SP