MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

# Reconstructing Spectral Vectors with Uncertain Spectrographic Masks for Robust Speech Recognition

Bhiksha Raj, Rita Singh

TR2005-160 November 2005

**Abstract**

Missing-feature methods improve automatic recognition of noisy speech by removing unreliable noise corrupted spectrographic components from the signal. Recognition is performed either by modifying the recognizer to work from incomplete spectra, or by estimating the missing components to reconstruct complete spectra. While the former approach performs optimal classification with incomplete spectrograms, the latter permits recognition with cepstral features derived from reconstructed spectra. Traditionally, spectral components are considered unequivocally reliable or unreliable. Research has shown that the use of soft masks that provide a probability of reliability to spectral components instead can improve the performance of missing feature mehtods that modify the recognizer. However, soft masks have not been employed by methods that reconstruct the spectrogram. In this paper we present a new MMSE algorithm for spectrogram reconstruction. Experiments show that the use of soft masks results in significantly improved performance as compared to reconstruction methods that use binary masks.

*IEEE Automatic Speech Recognition and Understanding Workshop*

# RECONSTRUCTING SPECTRAL VECTORS WITH UNCERTAIN SPECTROGRAPHIC MASKS FOR ROBUST SPEECH RECOGNITION

*Bhiksha Raj*

Mitsubishi Electric Research Labs
Cambridge, MA, USA

*Rita Singh*

Haikya Corp.
Watertown, MA, USA

## ABSTRACT

Missing-feature methods improve automatic recognition of noisy speech by removing unreliable noise corrupted spectrographic components from the signal. Recognition is performed either by modifying the recognizer to work from incomplete spectra, or by estimating the missing components to reconstruct complete spectra. While the former approach performs optimal classification with incomplete spectrograms, the latter permits recognition with cepstral features derived from reconstructed spectra. Traditionally, spectral components are considered unequivocally reliable or unreliable. Research has shown that the use of *soft masks* that provide a probability of reliability to spectral components instead can improve the performance of missing feature methods that modify the recognizer. However, soft masks have not been employed by methods that *reconstruct* the spectrogram. In this paper we present a new MMSE algorithm for spectrogram reconstruction. Experiments show that the use of soft masks results in significantly improved performance as compared to reconstruction methods that use binary masks.

## 1. INTRODUCTION

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

The actual recognition of the noisy speech can be performed in one of two ways. Most commonly, the recognizer, usually an HMM-based recognizer, is itself modified to work from incomplete spectrographic information (e.g. [1], [2]). In these approaches, the manner in which the recognition system computes likelihoods of classes or states is modified to account for the unreliability of some of the spectrographic components of the incoming speech. Unreliable components are marginalized out of the class (or state output) densities prior to computing likelihoods, conditioned on any bounds on the true value of the components that may be derived from the observed unreliable values. We refer to these approaches as "classifier compensation" methods since compensation for unreliable data is performed within the classifier (or recognizer).

In principle, classifier-compensation methods perform theoretically optimal classification and can therefore be expected to perform very well. However, in these methods the recognizer must explicitly model the distribution of the spectrographic features (typically log Mel spectra). Unfortunately, spectrographic features (such as log spectra) are suboptimal recognition; significantly superior recognition can be obtained with cepstral coefficients derived through linear transformations of the log spectra.

As an alternative approach, we have previously proposed the use of missing-feature methods that provide robust recognition through *feature compensation* [3]. These methods modify the incoming spectrographic features rather than the manner in which recognition is performed. Unreliable spectral components are erased and reconstructed using statistical information derived from clean speech and the remaining reliable components. This results in a set of complete log spectral vectors from which standard cepstral coefficients can be derived. Although this approach is not theoretically optimal, this disadvantage is often overcome by the improved recognition achieved in the cepstral domain.

In all cases, missing-feature methods depend critically on the accurate determination of the "spectrographic masks" that identify unreliable spectrographic components. Estimation of spectrographic masks, however, is a difficult task, since the very notion of "unreliability" in spectral components is not clearly defined. The reliability of spectral components is usually assumed to be indicated by their signal to noise ratio (SNR): components with an SNR of 0db or less are assumed to be unreliable (in reality the optimal SNR threshold for tagging unreliable components depends on the actual missing feature method employed [4]). However, it

is difficult to measure the SNR of spectral components of noisy speech, particularly when the corrupting noise is non-stationary or transient. Consequently, it is difficult to be certain if any particular spectral component is reliable. Thus, any technique that makes binary estimates of the reliability of spectral components will make mistakes, identifying reliable components as unreliable and vice versa. Such errors affect the performance of missing feature methods adversely.

In [5] Barker et. al. describe a classifier-compensation missing feature method that can utilize *soft masks*, i.e. spectrographic masks that associate a *probability of reliability* with each spectrographic component instead of tagging them in a binary manner as unequivocally reliable or unreliable. Soft masks avoid the pitfalls of erroneous binary identification of unreliable spectral components by merely associating with them a measure of confidence in their reliability. In [5] the authors show that the performance of missing feature methods can be greatly improved through the use of such soft masks.

The use of soft masks has thus far been restricted to classifier compensation missing feature methods. In this paper we present a minimum mean squared estimator (MMSE) based feature-compensation algorithm that utilizes soft masks to reconstruct spectral vectors. As in the work of Barker et. al., a real-valued number between 0 and 1 is associated with each spectrographic component. Since potentially every component of every spectral vector now has a non-zero probability of being unreliable, all spectral components must be estimated (rather than just the subset of components that have been tagged as unreliable). In order to do so, the observed noisy spectrogram is modelled as the output of a noisy channel where every component is either let through unchanged with some probability, or modified by an additive noise. The input to the channel are the log spectral vectors of clean speech. The probability distribution of these vectors is modelled by a mixture Gaussian density. The MMSE algorithm attempts to estimate the value of the input to the channel, given the noisy output that is observed, the probability with which the channel corrupts the input (which is given by the soft mask) and a simple assumed model for the distribution of the corrupting noise. Once all components of all log spectral vectors are estimated, cepstral vectors are derived from them and used for recognition.

Experiments conducted on a digits database corrupted to different degrees by four varieties of noise show that the proposed soft-mask-based spectral reconstruction method can result in significantly improved recognition over previously proposed feature compensation algorithms that use binary masks. The recognition performance obtained with cepstra derived from reconstructed spectral vectors is also found to be superior to that obtained with classifier compensation missing feature methods, as in previous studies at all evaluated SNRs.

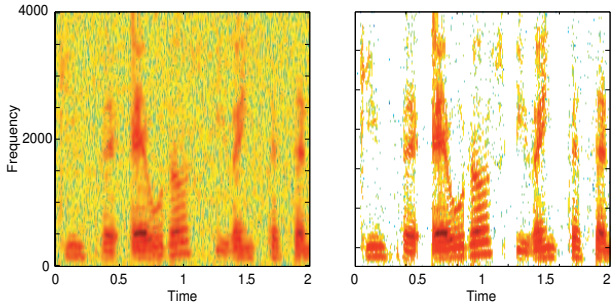



**Fig. 1**. Spectrogram of a speech signal corrupted to 10db by white noise. In the right panel all components with SNR less than 0db have been erased.

The rest of the paper is arranged as follows: In Section 2 we briefly describe the principles behind missing-feature methods. In Sections 3 and we briefly review some relevant current missing feature methods and the techniques used to obtain binary and soft masks for noisy data. In Section 4 we describe the proposed soft mask based MMSE algorithm. In Section 5 we describe our experimental results and in Section 6 we present our conclusions.

## 2. MODELLING NOISY SPEECH WITH INCOMPLETE SPECTROGRAMS

The speech signal is a highly non-stationary signal with spectral characteristics that vary both with time and frequency. When a speech signal is corrupted by noise, some of its time-frequency components are affected to a greater degree than others. The components of any time-frequency representation of the signal such as a spectrogram will therefore exhibit varying SNRs. High-SNR components from such a representation chiefly represent the characteristics of the speech and provide reliable information about the underlying phonetic content of the signal. Low-SNR components, on the other hand, also represent the characteristics of the noise and cannot be relied upon to represent the underlying speech. The only reliable measurement that can be derived from them is an *upper bound* on the true (noise-free) value of the components, if the noise is assumed to be additive.

Missing feature methods are based on the premise that speech recognition accuracy on noise-corrupted speech can be improved greatly if the evidence required for recognition were derived primarily from the reliable high-SNR components of time-frequency representations of the signal, deriving only minimal bounding information from the unreliable low-SNR components. This is illustrated by Figure 1. The left panel shows the spectrogram of a speech signal that has been corrupted to 10dB by white noise. In the figure in the right panel all spectrographic components with a local SNR

less than 0dB have been deemed unreliable and have been erased. Recognition must now be performed using only the incomplete data represented in the right panel.

The spectrographic representations used in missing feature methods are typically sequences of Mel-scaled log spectral vectors derived from the speech signal. Reliable and unreliable components are identified on these vectors.

## 3. REVIEW OF CURRENT TECHNIQUES

In this section we briefly review some current missing feature algorithms, as well as some methods for estimating spectrographic masks that have been used for comparative evaluations in Section 5.

### 3.1. Missing-Feature Algorithms

Recognition with incomplete spectrograms such as the one shown in Figure 1 can be done in one of two ways. Classifier compensation methods modify the recognizer to perform recognition with only the reliable (visible) regions of the spectrogram. Feature compensation methods reconstruct complete spectrograms by estimating the true value of the components in the unreliable (blanked out) regions and perform recognition with features derived from the complete spectrogram. We describe some fundamental missing feature algorithms of both varieties below. Bounded marginalization and soft mask based marginalization are classifier modification methods, while cluster based reconstruction is a feature compensation method.

#### *3.1.1. Bounded Marginalization*

In bounded marginalization [2], the unreliable components of a log-spectral vector are integrated out of the distribution of a class, constrained by upper and lower bounds on the true values of these components implicit in their observed values. For HMM-based recognizers that model state output densities as mixtures of Gaussians with diagonal covariance matrices, this results in the following modification in the computation of the contribution of the $d^{\text{th}}$ dimension of any log spectral vector to the state output densities of the HMM:

$$P(x_d|k,s,\theta_d) = \begin{cases} \int_{L_d}^{x_d} P(x_d|k,s)dx_d & \theta_d = 0 \\ P(x_d|k,s) & \theta_d = 1 \end{cases} \quad (1)$$

where $x_d$ is the $d^{\text{th}}$ component of a log-spectral vector $x$, $P(x_d|k,s)$ is the $d^{\text{th}}$ component of $k^{\text{th}}$ Gaussian in the mixture Gaussian density for state $s$, and $L_d$ is an empirically derived lower bound on the true value of $x_d$. $\theta_d$ is a binary tag, obtained from the spectrographic mask for the signal, that takes the value 1 when $x_d$ is reliable, and 0 when it is not. $P(x_d|k,s,\theta_d)$ is used in lieu of $P(x_d|k,s)$ to perform recognition.

#### *3.1.2. Soft Mask based Marginalization*

Soft mask based marginalization [5] is similar to bounded marginalization with the difference that the mask variable $\theta_d$ now represents the probability that $x_d$ is reliable and takes real values between 0 and 1. The state output density component described by Equation 1 gets modified to:

$$P(x_d|k,s,\theta_d) = \theta_d P(x_d|k,s) + (1-\theta_d)\frac{\int_{L_d}^{x_d} P(x_d|k,s)dx_d}{x_d - L_d} \quad (2)$$

Equation 2 can be derived from a model assumption that is also used by the MMSE algorithm presented in this paper and is described in Section 4.

#### *3.1.3. Cluster-Based Reconstruction*

Cluster-based reconstruction [3] is a feature compensation method that reconstructs complete log spectral vectors from noisy vectors with unreliable components (identified thusly by a binary mask). Here, the distribution of the log-spectral vectors of clean speech is modelled by a mixture Gaussian:

$$P(x) = \sum_k c_k \mathcal{N}(x;\mu_k,\Omega_k) \quad (3)$$

$\mathcal{N}(x;\mu_k,\Omega_k)$ represents a Gaussian with mean $\mu_k$ and variance $\Omega_k$. $c_k$ is the mixture weight of the $k^{\text{th}}$ Gaussian. To estimate the true value of unreliable components, the *a posteriori* probability of each Gaussian is first computed:

$$P(k|x) = Z\prod_d P(x_d;k,\theta_d) \quad (4)$$

where $x_d$ is the $d^{\text{th}}$ component of $x$. $P(x_d;k,\theta_d)$ is computed analogously to Equation 1 and $Z$ is a normalizing constant. The estimated value for any unreliable component $x_d$ is obtained as a linear combination of Gaussian-dependent maximum *a posteriori* (MAP) estimates:

$$\hat{x_d} = \sum_k P(k|x) MAP(x_d|x,\mu_k,\Omega_k) \quad (5)$$

The result of the operation is a complete spectrogram where all unreliable components have been reestimated. Cepstra derived from the spectrograms are used for recognition.

### 3.2. Estimating Spectrographic Masks

The most difficult component of missing-feature approaches is the estimation of spectrographic masks. In this paper we have used the MaxVQ algorithm [6] to generate binary spectrographic masks and the soft mask algorithm of [7] to generate probabilistic soft masks. Both algorithms assume that the distribution of the corrupting noise is known - an assumption that was valid for the experiments reported in this paper. We outline the two algorithms below.

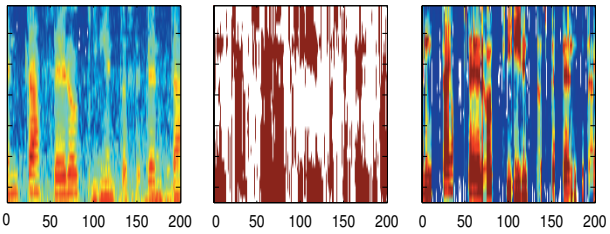



**Fig. 2**. a) Mel spectrogram of a noisy signal b) Binary mask from MaxVQ c) Soft mask

### *3.2.1. The Max-VQ algorithm*

The MaxVQ algorithm models the distributions of the log spectral vectors of speech, $x$ and noise, $n$ as mixtures of Gaussians with diagonal covariance matrices. The probability density of the log spectral vectors of the noisy speech, $y$, is assumed to have the following form:

$$P(y) = \sum_k \sum_j c_k^x c_j^n \mathcal{N}(y; \max(\mu_k^x, \mu_j^n), \Omega) \qquad (6)$$

where $c_k^x$ and $c_j^n$ are the mixture weights of the $k^{\text{th}}$ and $j^{\text{th}}$ Gaussians respectively from the mixture Gaussian distributions of speech and noise, and $\mu_k^x$ and $\mu_j^n$ are the corresponding means. The $\max(\cdot)$ is a component-by-component operator.

In order to find the spectrographic mask for any noisy vector $y$, the most likely combination $(k_{max}, j_{max})$ of Gaussians from the distributions of speech and noise is determined. The binary mask for any component $x_d$ is obtained as $\theta_d = 1$ if $\mu_{k_{max},d}^x > \mu_{j_{max},d}^n$; 0 else. Figure 2b shows an example of a spectrographic mask derived by MaxVQ.

### *3.2.2. Soft Mask Estimation*

The soft mask estimation algorithm also models the distributions of the log spectra of speech and noise by mixture Gaussians. A noisy log spectral vector $y$ is assumed to be related to the log spectra of the underlying speech and noise as $y = \max(x, n)$. The soft mask for $y_d$, the $d^{\text{th}}$ component of a noisy log spectral vector $y$ is obtained as:

$$\theta_d = \sum_k \sum_j \frac{P(k,j|y) P_x(y_d|k) C_n(y_d|j)}{C_x(y_d|k) P_n(y_d|j) + P_x(y_d|k) C_n(y_d|j)} \qquad (7)$$

where $P_x(y_d|k)$ and $C_x(y_d|k)$ represent the Gaussian density value and Cumulative probability at $y_d$ of the $d^{\text{th}}$ dimension of the $k^{\text{th}}$ Gaussian for the speech and $P_n(y_d|j)$ and $C_n(y_d|j)$ are similar terms for the distribution of the noise. Figure 2c shows an example of a soft mask derived in this manner. Note that in SNR terms, this mask represents the probability that any spectrographic component has SNR greater than 0.

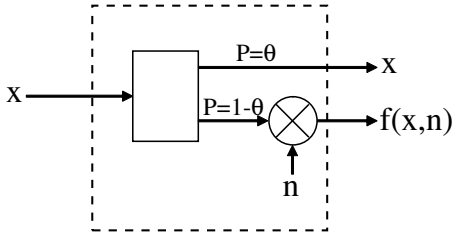



**Fig. 3**. Noisy channel model for soft masks

## 4. MINIMUM MEAN SQUARE ESTIMATION OF SPECTRAL COMPONENTS FROM SOFT MASKS

In this section we describe the proposed MMSE algorithm for reconstructing spectral vectors from soft masks. The algorithm models the log spectral vectors of noisy speech as the output of a noisy channel. Figure 3 illustrates the model. The input to the noisy channel are the $x_d$ components of the the log spectral vectors of clean speech and the output are the components of the noisy log-spectral vector.

The operation of the channel may be described as follows. As before, we represent the log spectral vectors of clean speech by $x$. The probability distribution of the log spectral vectors of clean speech is assumed to be a mixture Gaussian with diagonal covariance matrices:

$$P(x) = \sum_k c_k^x \mathcal{N}(x; \overset{x}{k}, \Omega_k^x) \qquad (8)$$

In order to generate an output $y$, a Gaussian is drawn from the mixture, a vector is drawn randomly from the Gaussian, and the components of the vector are transmitted through the channel. A separate channel is assumed for each dimension of the log-spectral vector. The channel transmits the input unchanged to the output with a probability $\theta_d$. With probability $1 - \theta_d$ it corrupts the input during the transmission. To corrupt the input it randomly draws a noise sample $n_d$ from a distribution $P_n(n_d)$, and combines it with the input $x_d$ through a function $f(\cdot)$ such that $f(x_d, n_d) \geq x_d$. The noise-corrupted output of the channel $y_d = f(x_d, n_d)$. The parameter $\theta_d$ which represents the value of the soft mask, and the distribution of the noise $P_n(n_d)$ are assumed to be different for each channel. The distribution of the output of the channel for the $d^{\text{th}}$ dimension of the log spectral vectors, given that the input vector has been drawn from the $k^{\text{th}}$ Gaussian is given by:

$$\begin{aligned} P_y(y_d|k) \quad = \quad & \theta_d P_x(y_d|k) + \\ & (1-\theta_d) \int_{-\infty}^{\infty} P_x(z_d|k) P_n(f^{-1}(y_d, z_d)) dz_d \end{aligned} \qquad (9)$$

where $f^{-1}(y_d, z_d)$ is the inverse function of $f(\cdot)$ that computes the set of all $n_d$ values such that $f(z_d, n_d) = y_d$.

$P_x(y_d|k)$ is the Gaussian $\mathcal{N}(y_d; {}^{x}_{k,d}, \Omega^x_{k,d})$. We assume that $P_n(f^{-1}(y_d, z_d)$ is a uniform probability distribution between $f^{-1}(y_d, y_d)$ and $f^{-1}(y_d, L_d)$, where $L_d$ is some known constant (typically set to the lowest possible value of $x_d$). Using these values, we now get

$$P_y(y_d|k) = \theta_d P_x(y_d|k) + \frac{1-\theta_d}{H_d} \int_{L_d}^{y_d} P_x(z_d|k) dz_d \quad (10)$$

where $H_d$ is a normalizing constant[1]. Note that Equation 10 is identical to Equation 2 of Section 3.1.3. The overall probability distribution of $y$ is given by

$$P_y(y) = \sum_k c^x_k \prod_d P_y(y_d|k) \quad (11)$$

The *a posteriori* probability, given $y$, of the $k^{\text{th}}$ Gaussian is

$$P(k|y) = \frac{c_k \prod_d P_y(y_d|k)}{\sum_j c_j \prod_d P_y(y_d|j)} \quad (12)$$

It can now be shown that the *a posteriori* probability of $x_d$ given $y$ and Gaussian index $k$ is given by

$$P_x(x_d|y,k) = \begin{cases} \theta_d \delta_{x_d}(y_d) & +(1-\theta_d)\frac{P_x(x_d|k)}{C_x(y_d|k) - C_x(L_d|k)} \\ & \text{if } L_d \le x_d \le y_d \\ 0 & \text{else} \end{cases} \quad (13)$$

where $\delta_{x_d}(y_d)$ is a Kronecker delta function centered at $y_d$ and, as before, $C_x(y_d|k)$ represents the cumulative probability at $y_d$ of the $d^{\text{th}}$ dimension of the $k^{\text{th}}$ Gaussian. The overall *a posteriori* probability of $x_d$ is given by

$$P_x(x_d|y) = \sum_k P(k|y) P_x(x_d|y,k) \quad (14)$$

The MMSE estimate of $x_d$ is simply the expected value of $x_d$, given the observed vector $y$. To obtain the MMSE estimate of $x_d$ we draw upon the following identity:

$$\int_{-\infty}^{a} x\mathcal{N}(x;\mu,\sigma)dx = \mu \int_{-\infty}^{a} \mathcal{N}(x;\mu,\sigma)dx - \sigma\mathcal{N}(a;\mu,\sigma) \quad (15)$$

Combining Equations 15, 14 and 13 we get the following MMSE estimate for $x_d$

$$\hat{x}_d = \theta_d y_d + (1-\theta_d) \sum_k P(k|y) \cdot \left( \mu^x_{k,d} - \Omega^x_{k,d} \frac{P_x(y_d|k) - P_x(L_d|k)}{C_x(y_d|k) - C_x(L_d|k)} \right) \quad (16)$$

The MMSE estimates $\hat{x}_d$ are arranged into a vector $\hat{x}$ that is used to compute cepstra that are used for recognition.

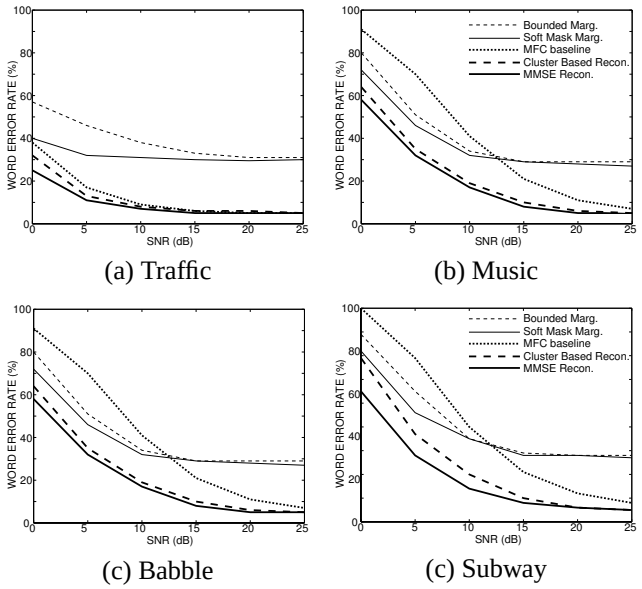
[1] If we assume $f(x,y) = x + y$, $H_d = y_d - L_d$

## 5. EXPERIMENTAL EVALUATION

The proposed soft-mask-based MMSE feature compensation algorithm was evaluated on a Spanish telephone speech database provided by Telefonicá Investigación y Desarrollo (TID), using the CMU Sphinx-3 speech recognition system. Continuous density 8 Gaussian/state HMMs with 500 tied states were trained from 3500 utterances of clean telephone recordings. The test data consisted of telephone recordings corrupted to various SNRs by traffic noise, music, babble recorded in a bar, and noise recordings from a subway. A total of 1700 test utterances were used in each case. The distribution of the log spectral vectors of clean speech was modelled by a 512 component mixture Gaussian density, the parameters of which were trained from the 3500 utterance training corpus. The distributions for the noises were modelled as mixtures of 256 Gaussians, the parameters of which were learned from training examples of the noises. A separate distribution was learned for each noise.

In all cases it was assumed that the type of noise affecting the speech was known. Binary spectrographic masks were estimated for all noise-corrupted utterances using the MaxVQ algorithm described in Section 3.2.1. Soft masks were estimated using the soft mask estimation algorithm described in Section 3.2.2. Two separate recognition experiments were conducted. In the first, acoustic models were trained with the log-spectral vectors of clean speech. No difference or double difference features were employed. No mean normalization of the training data was performed. Recognition was performed using the classifier modification methods presented in Section 3.1, namely bounded marginalization and soft mask based marginalization. Note that it is difficult to employ mean normalization with classifier modification methods. The benefit from difference and double difference features is also greatly reduced since they can have upto twice or four times as many unreliable components as the basic log spectral vectors themselves.

In the second experiment the recognizer was trained with cepstral vectors from clean speech. Mean normalization was performed. Difference and double difference features were also employed. For the noisy test data complete log spectral vectors were constructed using the cluster-based algorithm of Section 3.1.3 and the proposed MMSE algorithm. Since the former utilizes binary spectrographic masks while the latter employs soft masks, the difference in performance between the two shows the improvements to be obtained from the use of soft masks. The four panels in Figure 4 show the recognition performance obtained on speech corrupted by each of the four varieties of noise. In each panel, baseline recognition with cepstra, recognition obtained by bounded marginalization of log spectra using hard masks, the performance obtained by soft-mask-based marginalization of log spectra, and that obtained with cepstra derived




(a) Traffic (b) Music

(c) Babble (c) Subway

**Fig. 4**. Recognition error vs. SNR on speech corrupted by a) traffic noise b) music c) babble and d) subway noise

from spectra reconstructed using cluster-based reconstruction (using hard masks) and the proposed soft-mask-based MMSE technique are all shown.

## 6. CONCLUSIONS

We observe from Figure 4 that the proposed soft mask based MMSE algorithm results in significantly improved recognition over cluster based reconstruction, which uses binary masks, particularly for traffic and subway noises where it results in a 25% relative improvement at 0dB. In general, we also observe that soft-mask based methods are superior to missing feature methods that utilize binary spectrographic masks.

We also observe that in these experiments, feature compensation algorithms combined with recognition in the cepstral domain significantly outperform classifier compensation methods that work in the log spectral domain, in spite of the fact that the latter are theoretically optimal in the log spectral domain. This is consistent with results published previously, e.g. [3]. The proposed MMSE method, which is a feature compensation method, is observed to result in the best recognition of all four methods evaluated. On the other hand, the difference between classifier compensation and feature compensation methods is observed to decrease at very low SNRs. In some of the experiments, classifier compensation methods working from log spectral vectors actually result in greatly improved performance over baseline cepstra-based recognition at SNRs below 15dB.

We wish to emphasize here that the purpose of this paper is to present a spectrogram reconstruction technique that works from soft masks. The comparison between feature compensation and classifier compensation methods is only a peripheral, but related topic. It is likely that incorporation of difference features and percentile-based normalization (proposed in [8] as a substitute for mean normalization) will improve the performance of our implementation of classifier compensation methods and reduce the difference in performance obtained by the two classes of missing feature approaches.

## 7. REFERENCES

[1] M.P. Cooke, A. Morris, and P.D. Green, "Missing data techniques for robust speech recognition," in *IEEE Conf. on Acoustics, Speech and Signal Processing*, 1997.

[2] M.P. Cooke, P.D. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2000.

[3] B. Raj, R.M. Stern, and M.L. Seltzer, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[4] B. Raj, *Reconstruction of Incomplete Spectrograms for Robust Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, April 2000.

[5] J. Barker, L. Josifovski, M.P. Cooke, and P.D. Greene, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Intl. Conf. on Speech and Language Processing*, 2000.

[6] S. Roweis, "One microphone speaker separation," in *EUROSPEECH*, 2003.

[7] A. Reddy and B. Raj, "Soft-mask estimation for single channel speaker separation," in *ISCA ITRW on Statistical and Perceptual Audio Processing*, 2004.

[8] K.J. Palomaki, G.J. Brown, and J. Barker, "Techniques for handling convolutional distortion with missing data automatic speech recognition," *Speech Communication*, vol. 43, pp. 123–142, 2004.