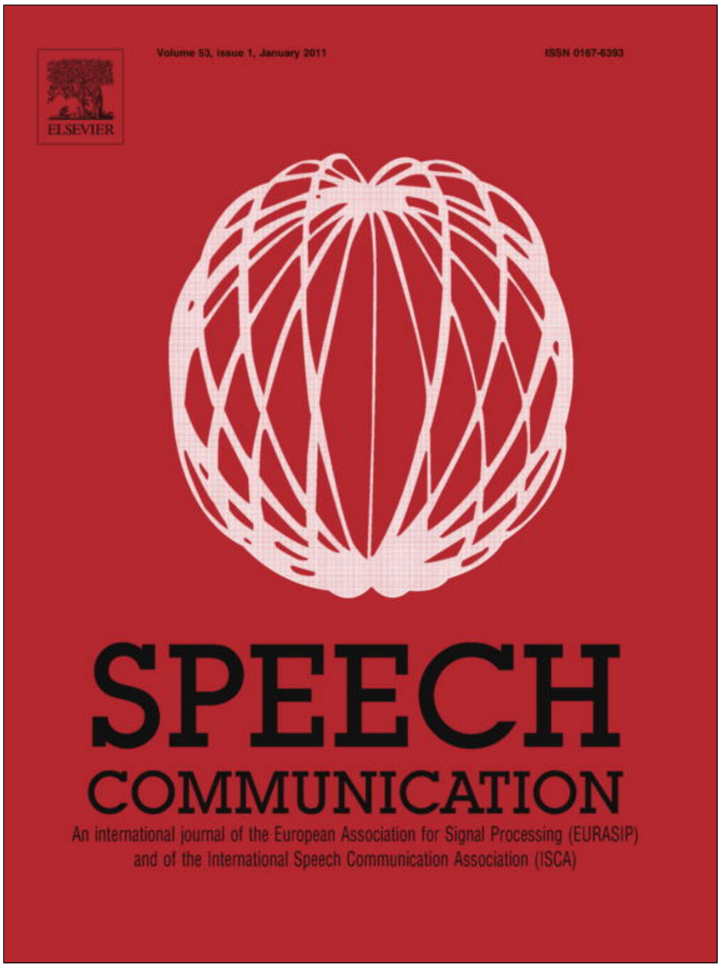
Volume 53, Issue 1, January 2011
ISSN 0167-6393
ELSEVIER
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP)
and of the International Speech Communication Association (ISCA)

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 1–11
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise

Wooil Kim [a,*], Richard M. Stern [b]

[a] Center for Robust Speech Systems (CRSS), Erik Jonsson School of Engineering and Computer Science, Department of Electrical Engineering, University of Texas at Dallas, 2601 N. Floyd Road, EC33, Richardson, TX 75080-1407, USA
[b] Department of Electrical and Computer Engineering and Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Received 30 June 2009; received in revised form 6 July 2010; accepted 5 August 2010

## Abstract

"Missing-feature" techniques to improve speech recognition accuracy are based on the blind determination of which cells in a spectrogram-like display of speech are corrupted by the effects of noise or other types of disturbance (and hence are "missing"). In this paper we present three new approaches that improve the speech recognition accuracy obtained using missing-feature techniques. It had been found in previous studies (e.g. Seltzer et al., 2004) that Bayesian approaches to missing-feature classification are effective in ameliorating the effects of various types of additive noise. While Seltzer et al. primarily used white noise for training their Bayesian classifier, we have found that this is not the best type of training signal when noise with greater spectral and/or temporal variation is encountered in the testing environment. The first innovation introduced in this paper, referred to as frequency-dependent classification, involves independent classification in each of the various frequency bands in which the incoming speech is analyzed based on parallel sets of frequency-dependent features. The second innovation, referred to as colored-noise generation using multi-band partitioning, involves the use of masking noises with artificially-introduced spectral and temporal variation in training the Bayesian classifier used to determine which spectro-temporal components of incoming speech are corrupted by noise in unknown testing environments. The third innovation consists of an adaptive method to estimate the *a priori* values of the mask classifier that determines whether a particular time-frequency segment of the test data should be considered to be reliable or not. It is shown that these innovations provide improved speech recognition accuracy on a small vocabulary test when missing-feature restoration is applied to incoming speech that is corrupted by additive noise of an unknown nature, especially at lower signal-to-noise ratios.

*Keywords:* Robust speech recognition; Missing-feature reconstruction; Frequency-dependent mask classification; Colored-noise masker generation; Multi-band partition method

## 1. Introduction

Differences in acoustic environment between the conditions under which an automatic speech recognition (ASR) system is trained and deployed are primary factors underlying degradation in speech recognition accuracy, including the presence of background noise. Various approaches for minimizing these differences and for maximizing speech recognition accuracy have been developed over the last several decades, and these algorithms have achieved reasonable success in the presence of stationary noise. Nevertheless, these approaches are still vulnerable to the effects of time-varying noise such as background music, since most of them are primarily based on the estimation of corrupting noise components. In general it is very difficult to estimate the statistical characteristics of unknown background noise that is time varying, and the presence of such noise greatly complicates environmental compensation efforts (e.g. Singh et al., 2002a,b).

* Corresponding author. Tel.: +1 972 883 4388; fax: +1 972 883 2710.
*E-mail address:* wikim@utdallas.edu (W. Kim).
*URL:* http://crss.utdallas.edu (W. Kim).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.005

Missing-feature methods have been more effective in coping with the effects of non-stationary noise on speech recognition accuracy. These methods depend mostly on the characteristics of speech that are resistant to noise, rather than on the characteristics of the noise itself. In principle, missing-feature techniques should enable an improvement in accuracy that is independent of the specific nature of the masking background noise, even when the noise is transient in nature (Cooke et al., 1997; Lippmann and Carlson, 1997; Cooke et al., 2001; Raj et al., 2004; Seltzer et al., 2004; Raj and Stern, 2005).

Missing-feature methods generally consist of two steps. The first step is the estimation of a "mask" which determines which parts of a spectro-temporal representation of noisy input speech are considered to be unreliable. Most of the initial mask classification methods proposed were based on the estimation of quasi-stationary background noise of a known type. These methods are not useful in estimating unknown or time-varying noise processes (Barker et al., 2000; Renevey and Drygajlo, 2001). Seltzer et al. (2004) have proposed a Bayesian classifier with the goal of environment-independent mask classification. Jancovic et al. (2003) have suggested a method to evaluate the reliability of a particular band using the likelihood computed from a Hidden Markov Model (HMM). Harding et al. (2005), Srinivasan et al. (2006), and Park and Stern (2009) have described methods that exploit spatial information from multiple channels of incoming speech to estimate the masks for the missing-feature algorithm. In addition, Raj and Singh (2005) proposed a method for creating "soft" masks called the Max-VQ algorithm, which still relies on estimation of the noise.

The second step concerns the determination of how to bypass the unreliable (or "missing") regions that are identified in the first step, or reconstruct "reliable" features from them. Early methods involved modifying the decoding procedures of the speech recognizer to compute the output probabilities associated with the incomplete speech. These approaches would either replace unreliable spectro-temporal regions by estimating their values given the state of the HMM or by computing the marginal output probability, relying only on the reliable regions while bypassing the unreliable regions (Cooke et al., 2001; Josifovski et al., 1999). Raj et al. (2004) have proposed two types of feature-based reconstruction methods. Our work is based on one of these approaches, cluster-based missing-feature reconstruction, which will be described in a later section.

The work described in the present paper focuses on the first step, mask classification. Seltzer et al. (2004) described the development of a Bayesian classifier for mask classification. They mostly trained on speech corrupted by white noise for this work because it was noted empirically that classifiers trained in white noise tend to be somewhat more robust to changes in the acoustical environment. Nevertheless, we found in subsequent evaluations that the use of white noise for training the Bayesian classifier does not in fact provide the desired degree of environment independence in mask classification, for reasons that we believe are related to the inability of white noise to reflect spectral variations of realistic noise environments realistically over time and frequency. For this reason we have proposed a new training method that employs a combination of colored-noise samples, and we will demonstrate that the use of this method improves environmental robustness.

In Section 2 we review the Bayesian classifier for mask classification proposed by Seltzer et al. (2004), which is used as the baseline system for our experiments. In Section 3 we review the cluster-based reconstruction algorithm of Raj et al. (2004), which is employed for the missing-feature reconstruction method in our work. We generalize the problem of mask classification for missing-feature recognition in Section 4, modifying the classifier so that it operates in a "frequency-dependent" fashion. In Section 5 we introduce a new way to improve the environmental robustness of the classifier by training it using a particular type of colored-noise broadband noise. Finally, in Section 6 we summarize our findings.

## 2. Mask classification based on Bayesian classification

The missing-feature approach requires that we determine a "mask" which classifies the spectral components of each frame into reliable and unreliable (or "missing") regions for missing-feature reconstruction. Reliable regions are defined as the spectro-temporal components of incoming speech in which the speech components remain undistorted by the corrupting background noise. In the unreliable regions, the noise components are intense enough to distort the representation of speech to the extent that the representation is no longer useful for speech recognition.

Seltzer et al. (2004) have proposed a Bayesian classifier for mask classification that makes no assumption about the nature of the corrupting background noise. Their method employs measures of speech attributes which assess the degree of corruption by noise while remaining relative robust to the nature of the background noise. These features include the following:

- The comb filter ratio (CFR), which measures the ratio of energy at frequencies that are harmonics (or integer multiples of the fundamental frequency) compared to the energy between these frequencies.
- The ratio of subband energy to full-band energy.
- The ratio of subband energy to the full-band/subband noise floor.
- Spectral flatness, which is characterized by the variance of the subband energy in a neighborhood of spectro-temporal locations.

Seltzer et al. (2004) have demonstrated that the CFR is a reliable predictor of noise-level in the signals, by showing its performance with voiced speech segments corrupted by white noise and background music at various SNRs.

The CFR is extracted only during voiced frames from which the pitch period for the comb filter can be obtained. Using the features above, acoustic models are estimated separately for voiced and for unvoiced speech segments, based on observed log-spectral values in each Mel-filter band. Prior probabilities and Gaussian mixture models (GMMs) are developed for each acoustic model. Seltzer et al. obtained these acoustic model parameters by training on a speech database that was artificially corrupted by noise, and the labels for reliable versus unreliable segments were developed using "Oracle" information[1] which contains perfect knowledge about the nature of the corrupting noise. Training for mask classification will be discussed further in Section 5. When new data are presented to the ASR system, the determination of whether a given spectral-temporal component is reliable or unreliable is based on its *a posteriori* probability in terms of these models.

## 3. Cluster-based missing-feature reconstruction

Our missing-feature reconstruction is an extension of the cluster-based method introduced by Raj et al. (2004). Using maximum *a posteriori* (MAP) estimation techniques, the unreliable components of speech representations are estimated using values of the reliable regions (as determined by the mask classification process), based on the known distributions of clean speech. The feature vectors are reconstructed in the log-spectral domain, and then converted to cepstral features for the actual speech recognition. The use of cepstral coefficients as feature vectors for speech recognition can provide better recognition accuracy than model compensation methods that restore log-spectral components (Raj et al., 2004).

Let $X$ represent the log spectra of clean speech, which are modeled by Gaussian-mixture densities with $K$ clusters. Each cluster has a mean vector and a full covariance matrix,

$$p(X) = \sum_{k=1}^{K} \omega_k \mathcal{N}(X; \boldsymbol{\mu}_{X,k}, \boldsymbol{\Sigma}_{X,k}). \tag{1}$$

Consider a noisy speech vector $Y$ and its underlying clean speech vector $X$ which has unreliable (i.e. "missing") components $X_u$ and reliable components $X_r$. The reliable components $X_r$ are identical to the corresponding observations $Y_r$. The cluster membership $k$ of the clean speech model is nominally determined by its *a posteriori* probability, which can be computed by integrating out the unreliable elements:

$$\begin{aligned}\hat{k} &= \arg\max_k \{P(k)p(X|k)\} \\ &= \arg\max_k \left\{ P(k) \int_{-\infty}^{Y_u} p(X|k) dX_u \right\},\end{aligned} \tag{2}$$

where $Y_u$ represents the observed values of the unreliable parts and is assumed to be greater than $X_u$ because it is corrupted by additive background noise. Finally, the unreliable parts $X_u$ are restored using bounded MAP estimation based on the observations in the reliable regions $X_r$, the model parameters of the cluster $\hat{k}$ as determined by (2), and the upper bound $Y_u$ as follows:

$$\widehat{X}_u = \arg\max_{X_u} \left\{ p(X_u | X_r, \boldsymbol{\mu}_{X,\hat{k}}, \boldsymbol{\Sigma}_{X,\hat{k}}, X_u \leqslant Y_u) \right\}, \tag{3}$$

where $\boldsymbol{\mu}_{X,\hat{k}}$ and $\boldsymbol{\Sigma}_{X,\hat{k}}$ respectively denote a mean vector and covariance matrix of the cluster $\hat{k}$ of the clean speech Gaussian mixture model.

## 4. A frequency-dependent Bayesian classifier for mask classification

In the work of Seltzer et al. (2004), the classifier for mask classification was trained using a speech database that was corrupted by white noise as described in Section 2. This classifier was then applied to other acoustical environments including factory noise and music noise in the background without any prior information about the test conditions. Recognition accuracy obtained using the factory-noise and musical maskers was comparable to the "matched" training condition in which the type of noise used both to train and evaluate the Bayesian mask was the same (i.e. white noise in this case).

In our extensions of this work, however, we found that the use of white noise to train the Bayesian mask classifier failed to provide good recognition accuracy in the presence of all types of corrupting noise. We believe that mask classifiers trained on white noise are suboptimal because of implicit frequency-to-frequency dependencies in the features that are used for mask classification. For example, the subband-energy-ratio features utilize the values of other subbands or a noise floor that is estimated from neighboring frames. In addition, the spectral-flatness feature directly exploits spectral variations around neighboring frames and frequency bands. In other words, we believe that the spectral variations across adjacent frames and frequency bands can influence the features obtained from a particular band. Therefore, in order to obtain environment-independent models for the Bayesian mask classifier, we must incorporate the spectral variations across frames and bands into the model of each band, which in effect simulates the occurrences of various kinds of noise conditions.

In an earlier study (Kim et al., 2005), we proposed a training method that uses combinations of colored noise for the purpose of generating an environment-independent model that can be used for mask classification. The effects of spectral variation across adjacent frames and frequency bands were incorporated by training the acoustic models for mask classification on speech databases that are corrupted by various random combinations of colored noise. The colored-noise samples are obtained by dividing the

[1] More details about the Oracle knowledge used in our study may be found in Section 4.2.

original frequency band into $M$ bands that increase in bandwidth as the center frequency increases according to the Mel scale.

In principle, any type of noise spectral profile can be generated by increasing $M$ to half the size of the discrete Fourier transform (DFT) and by employing a number of different signal-to-noise ratios (SNRs) across the frequency bins. Unfortunately, when the amount of training data is limited the frequency of occurrence of each type of colored noise decreases as the number of kinds of noise increases. From experimental observations we believe that this results in a failure to observe continued reduction in error rate as $M$, the number of frequency bands, is increased beyond a point. Increasing the size of the training database is not practical, because the database size must be increased in exponential proportion to $M$.

In this section, we present a frequency-dependent approach that addresses the problem of limited available training data which is caused by increasing the number of partitions for colored-noise generation (Kim and Stern, 2006). Within the framework of the frequency-dependent mask classifier, we need only to consider the various kinds of spectral events within each particular frequency band in order to simulate the various types of background noise considered. The mask classification scheme presented in the following section enables us to characterize the spectral patterns of a number of different background environments while using a relatively small number of combinations of colored noise. We begin with a description of the features used for frequency-dependent mask classification, followed by a discussion of the mask classification method. In Section 5 we discuss the method used to generate the stimuli that are used to train the mask classifiers.

### 4.1. Features used for frequency-dependent mask classification

The incoming signal is subdivided into 23 overlapping frequency bands that increase in bandwidth as the center frequency increases, following the dependence of bandwidth on center frequency that is used in conventional Mel-frequency cepstral analysis (Davis and Mermelstein, 1980). Mask classification in each frequency band is performed using a combination of some of the features used by Seltzer et al. (2004) with additional coefficients indicating spectral information, formulating 12 features for voiced speech segments and 11 features for unvoiced segments for each frequency band. (The fact that we happen to use 23 overlapping frequency bands as well as a total of 23 factors is merely a coincidence.)

#### 4.1.1. Subband cepstral coefficients

Cepstral coefficients provide an effective characterization of the short-time spectral envelope and can be used as features for mask classification, just as they are used for the speech recognition system itself. We develop "subband cepstral coefficients" by computing the discrete cosine transform (DCT) of the log magnitude spectrum in each of the 23 analysis bands (without the triangular weighting associated with conventional MFCC analysis). In each band, DCT coefficients 1 through 5 are used as features for mask classification.[2] In addition, we obtain five subband delta cepstral coefficients by computing the first difference of each of the subband cepstral coefficients in each Mel-frequency channel. We believe that the subband cepstral coefficients will represent the spectral envelope of each frequency band with a less correlation among the coefficients. We note that while classification within each frequency band is based on features that are specific to that band of frequencies, this classification takes place independently of the input to the other bands.

#### 4.1.2. Spectral flatness measure

The spectral flatness measure (SFM) indicates whether any tonal components are dominant in a given signal frame, and it has been used as a measure for determining which segments of an utterance are voiced or unvoiced (Johnston, 1988). The SFM can be calculated from the ratio of the geometric and arithmetic averages of spectral components as in Eq. (4):

$$\mathrm{SFM}(m) = \frac{\left\{\prod_n x_m(n)\right\}^{1/N_m}}{\frac{1}{N_m}\sum_n x_m(n)}, \tag{4}$$

where $x_m(n)$ indicates signal components in the (linear) spectral domain within the $m$th Mel-filter-bank, and $N_m$ denotes the number of these components. The SFM is expected to reflect the amount of contamination of each Mel-filter channel (i.e. Mel-frequency band) by background noise and it is computed from the log-spectral values in each frequency band.

#### 4.1.3. Comb filter ratio

The comb filter ratio (CFR) used by Seltzer et al. (2004) and described in Section 2 is used as the final feature for mask classification in each frequency band.

### 4.2. Frequency-dependent mask classification using Gaussian mixture models

In total, for each Mel-frequency band, the feature vector for voiced-speech frames consists of five subband cepstral coefficients, their corresponding first differences in time, one SFM coefficient, and one CFR coefficient, producing a 12-dimensional vector. Frames of unvoiced speech are represented by the same features excluding the CFR measure, producing an 11-dimensional feature vector for each frequency band.

The estimate-maximize (EM) method is used in conventional fashion to develop four Gaussian mixture models

[2] The smallest number of log-spectral components within each frequency band (i.e. Mel-filter-bank) is 5, when an analysis window of 256 samples is used.

(GMMs) in each band. Specifically, separate models are developed for spectral components that are deemed to be reliable and unreliable (i.e. "missing") for both voiced and unvoiced speech frames. Each model consists of a prior probability and a GMM of the distribution for the feature vectors, which enable us to calculate the posterior probability of the input feature vector representing each frame. The parameters of the models are obtained by a training procedure that used a multi-band type of colored noise to be described in the following section. For some of these experiments we make use of perfect "Oracle" knowledge that informs the system whether a particular time-frequency segment should be regarded as reliable or unreliable, for the purpose of identifying the extent to which recognition errors are mediated by mask classification errors as opposed to other causes. This Oracle mask of the $m$th frequency band at time $t$ is obtained by comparing a difference in log-spectral value between the original clean speech $s_m(t)$ (i.e. reference) and noise signal $\tilde{n}_m(t)$ estimated from input speech $x_m(t)$ to a threshold $\alpha$ as in Eq. (5):

$$O_m(t) = \begin{cases} 1 \text{ (reliable)}, & \text{if } s_m(t) - \tilde{n}_m(t) > \alpha, \\ 0 \text{ (unreliable)}, & \text{otherwise}, \end{cases} \quad 1 \leqslant m \leqslant M, \tag{5}$$

where $\tilde{n}_m(t) = \log(\exp(x_m(t)) - \exp(s_m(t)))$, if $x_m(t) > s_m(t)$. In our experiments, we used $-0.5$ for the threshold $\alpha$.

### 4.3. Experimental evaluation of frequency-dependent classification

#### 4.3.1. Experimental procedures

Our evaluations of the procedures described above were performed in the context of the Aurora 2.0 evaluation framework as developed by the European Language Resources Association (ELRA) (Hirsch and Pearce, 2000). The task is connected English-language digits consisting of eleven words, with each whole word represented by a continuous-density HMM with 16 states and three mixtures per state. In addition to the digits, two silence models representing normal silences and short pauses are employed. The feature extraction algorithm suggested by the European Telecommunication Standards Institute (ETSI) was employed for the experiments (ETSI, 2000). An analysis window of 25-ms duration is used with 10 ms between frames for speech that is sampled at 8 kHz. The computed magnitude spectrum is passed through a Mel-scaled filter-bank and 23 Mel-filter-bank outputs are transformed to 13 cepstral coefficients in the usual fashion (Davis and Mermelstein, 1980). After extracting the 13th-order cepstrum, discrete-time approximations to the first- and second-order time derivatives are included during the decoding procedure producing a final feature vector of 39 dimensions.

Following the procedures specified in the Aurora 2.0 evaluation for clean-condition training and multi-condition testing, the HMMs of the speech recognizer and the GMMs for cluster-based missing-feature reconstruction were trained using a database that contains 8440 utterances of clean speech. Rather than using the multi-condition testing database in Aurora 2.0, evaluation data were obtained by combining clean speech samples from Set A of the Aurora 2.0 testing database with four types of noise samples: white noise, car noise, speech babble, and background music. The white noise and car noise represent stationary noise conditions, and they were obtained from the NOISEX92 and Aurora 2.0 databases, respectively. Speech babble and background music represent non-stationary noise environments; they were obtained from the Aurora 2.0 database and the initial instrumental segments (before singing begins) of 10 Korean pop songs with varying degrees of intensity and speed (i.e. beat and tempo). The test database included speech samples that were corrupted by each of the four types of noise at five SNRs: 20, 15, 10, 5, and 0 dB. We obtained 1001 samples of degraded speech for each of the 20 noise conditions.

#### 4.3.2. Comparison of frequency-dependent classification with baseline performance

We now compare the performance of the frequency-dependent classification procedure described in this section with the baseline Bayesian classification procedure developed by Seltzer et al. (2004). The five features described in Section 2 were used for the baseline mask classification in voiced frames and the same features except for CFR were used for the unvoiced frames. The pitch information for CFR at every speech input was extracted using the histogram-based method described by Seltzer (2000). The features used for mask classification were modeled as Gaussian-mixture densities with 16 mixture components and diagonal covariance matrices.

Table 1 compares the speech recognition performance (Word Error Rate, or WER) of missing-feature reconstruction employing the frequency-dependent classification procedure described in this section with the baseline Bayesian classification procedure developed by Seltzer et al. (2004) for the four masker types, white noise, car noise, speech babble, and background music at all five SNR conditions. The performance obtained with no processing at all and with traditional spectral subtraction (Boll, 1979; Martin, 1994) are also presented. In each case, the mask classifier was trained using white noise that was presented at seven SNRs (clean, 20, 15, 10, 5, 0 and −5 dB), and the recognition was performed without attempting to estimate the SNR of each input sample. Different classifiers were applied for mask classification of voiced and unvoiced speech frames based on whether or not pitch was detected in each incoming speech frame. Fig. 1 compares the performance of the systems presented in Table 1, showing averages of WER over the five SNRs (0, 5, 10, 15, and 20 dB).

It can be seen that the missing-feature procedures considered here provide a very considerable improvement compared to the performance obtained with no processing at all. It also is clearly seen that with the exception of the

Table 1
Comparison of speech recognition performance in different background noise and SNR conditions using frequency-dependent classification with baseline classification and spectral subtraction (WER, %).

| | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB | Average |
|---|---|---|---|---|---|---|
| *White noise* | | | | | | |
| No processing | 79.93 | 48.58 | 25.62 | 11.60 | 5.55 | 34.26 |
| Spectral subtraction | 51.24 | 27.86 | 13.75 | 7.04 | 4.53 | 20.88 |
| Baseline classifier | 36.39 | 16.85 | 9.16 | 4.47 | 3.28 | 14.03 |
| Frequency-dependent classifier | 51.12 | 28.93 | 15.96 | 8.77 | 4.89 | 21.93 |
| *Car noise* | | | | | | |
| No processing | 88.07 | 63.91 | 27.71 | 8.38 | 2.92 | 38.20 |
| Spectral subtraction | 71.16 | 37.07 | 12.20 | 4.09 | 2.45 | 25.39 |
| Baseline classifier | 64.21 | 30.21 | 9.28 | 3.58 | 2.56 | 21.97 |
| Frequency-dependent classifier | 39.52 | 16.76 | 7.90 | 5.13 | 4.44 | 14.75 |
| *Speech babble* | | | | | | |
| No processing | 88.88 | 71.13 | 44.38 | 21.13 | 7.47 | 46.60 |
| Spectral subtraction | 85.91 | 63.33 | 38.78 | 19.62 | 8.46 | 43.22 |
| Baseline classifier | 73.91 | 49.33 | 25.06 | 12.39 | 5.35 | 33.21 |
| Frequency-dependent classifier | 46.70 | 23.52 | 10.55 | 6.17 | 3.81 | 18.15 |
| *Background music* | | | | | | |
| No processing | 74.27 | 51.34 | 28.11 | 12.19 | 4.84 | 34.15 |
| Spectral subtraction | 70.72 | 47.58 | 28.02 | 14.69 | 6.48 | 33.50 |
| Baseline classifier | 66.80 | 43.54 | 24.65 | 12.56 | 5.43 | 30.60 |
| Frequency-dependent classifier | 49.89 | 29.28 | 15.52 | 7.84 | 4.10 | 21.33 |

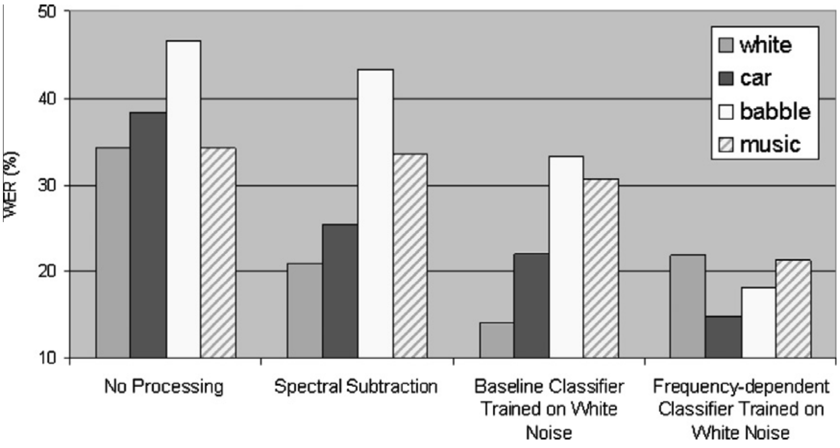

Fig. 1. Performance comparison of the systems in Table 1 as average WERs over the five SNRs.

white-noise masker the results obtained using the frequency-dependent Bayesian mask classification provide better overall performance, and (more significantly) that the variation in recognition accuracy with respect to masker type is sharply diminished. This suggests that each frequency band is independently trained by the spectral events of a given frequency region corresponding to the band in training of the mask classifier described in this paper, so that global similarity between training and testing conditions is not so important. The baseline classifier outperforms our frequency-dependent classification for white-noise maskers because in this case the training and testing conditions are perfectly matched using the Bayesian mask classifier developed by Seltzer et al. (2004).

While we are encouraged by these results, the results of pilot studies described in Kim et al. (2005) suggest that better results are obtained when the maskers used to train the mask classifier reflect the spectral patterns and variations that occur in the test conditions. In other words, while frequency-dependent training and testing provides substantial environmental independence, the white-noise maskers used to train the classifier in this section may not be effective in totally unknown environments as other types of maskers. We explore this issue in the next section.

## 5. Colored-noise generation using multi-band partitioning

In the previous section we described a frequency-dependent mask classifier that was trained using samples of a white-noise process. The experimental results in Section 4 and our pilot study suggest that performance could be further improved by training the mask classifier on maskers that exhibit a greater degree of local spectral and temporal variability than is provided by white noise. We believe that maskers of this sort resemble more closely the type of interference that is actually encountered in testing. In this section we describe a method for generating colored noise that can be used for training the frequency-dependent mask classifiers. The maskers that we propose exhibit greater spectral and temporal variability while at the same time are generic in nature and do not represent directly any particular testing environment.

### 5.1. Construction of artificial colored-noise signals for training mask classifiers

In our approach, each of the $M$ Mel-frequency bands of the original spectrum is divided into $N$ narrow parts (or "partitions"), with the partitions of each band unaffected by those of other bands. We refer to this approach as the "multi-band partition method" for generating colored noise (Kim and Stern, 2006).

In multi-band partitioning for colored-noise generation, each of the $N$ partitions of each Mel-frequency band may or may not contain noise components, resulting in $2^N$ spectral profiles that could be observed within each band. This partitioning of each band into further combinations of components is illustrated in Fig. 2. Because of the narrowness of the partitions, the colored-noise maskers used to train the system are generated by manipulating the frequency representations of the subbands directly, rather than through the use of bandpass filters. Specifically, spectral components of white noise are removed outside the frequency regions of interest of the spectra of the noise samples, and the resulting narrow-band restricted noise signal is obtained in the time domain through inverse Fourier transformation. The phase information of the original white noise sample is used for synthesizing the colored noise in a manner similar to that of conventional spectral subtraction.

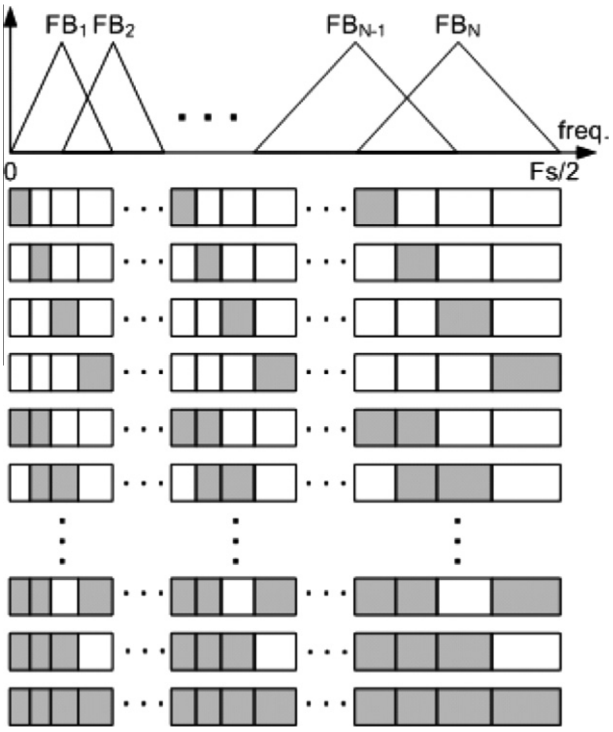

Fig. 2. Illustration of the multi-band partition method of generating artificial colored noise, with the case of four partitions per Mel-frequency band illustrated.

Fig. 3 is a spectro-temporal representation of the multi-band partition method for generating colored noise (again showing the case of $N = 4$). As the figure shows, each Mel-frequency band is split into four partitions of equal width along the Mel scale. The multi-band scheme with $N = 4$ has the effect of partitioning the entire range of frequencies into four times of the number of Mel-filter-banks $M$, (or $2^{4\times M}$ combinations) since the mask classifier corresponding to each band is trained independently. This means that we can simulate a greater number of spectral patterns with a relatively small number of combinations (i.e. $M \times 2^4$).

The $2^N$ combinations of colored noise in a Mel-frequency band that are used to train the mask classifier of the corresponding frequency index are generated by combining the narrow-band signals created by the $N$ partitions of each band. The exact combination is selected randomly for each of a sequence of successive time intervals (typically of duration 30, 60, or 300 ms in our experiments) if non-stationary noise is desired, or a single combination of partitions is selected for the entire utterance if stationary noise samples are desired. In the example shown in Fig. 3, the colored noise was generated by the multi-band partition method using $N = 4$, where each combination of colored noise is randomly selected every 30 ms. A noise-corrupted speech database for training the frequency-dependent mask classifier was produced by adding the colored-noise signals described above to clean speech at various SNRs. We used seven SNRs (clean, 20, 15, 10, 5, 0, and −5 dB) as a same manner as the model training in white noise presented, in Section 4.3.2.

### 5.2. Performance of frequency-dependent mask classification trained using multi-band colored noise

In this section we discuss the effectiveness of the artificial multi-band colored-noise signals described in Section 5.1, using the frequency-dependent mask-classification strategy discussed in Section 4. Fig. 4 presents the word error rate obtained for four different types of noise maskers, using the same 12 features for voiced speech and 11 features for unvoiced speech discussed previously. The figure shows how recognition accuracy depends on the number of frequency partitions that are used in generating the colored noise as in Fig. 2. Here the WER is an average value for all five SNR conditions (i.e. 0, 5, 10, 15, and 20 dB). Note that when there is only a single partition, the training masker becomes white noise, as was used in generating the training data for the mask classifiers in Fig. 1. As seen in

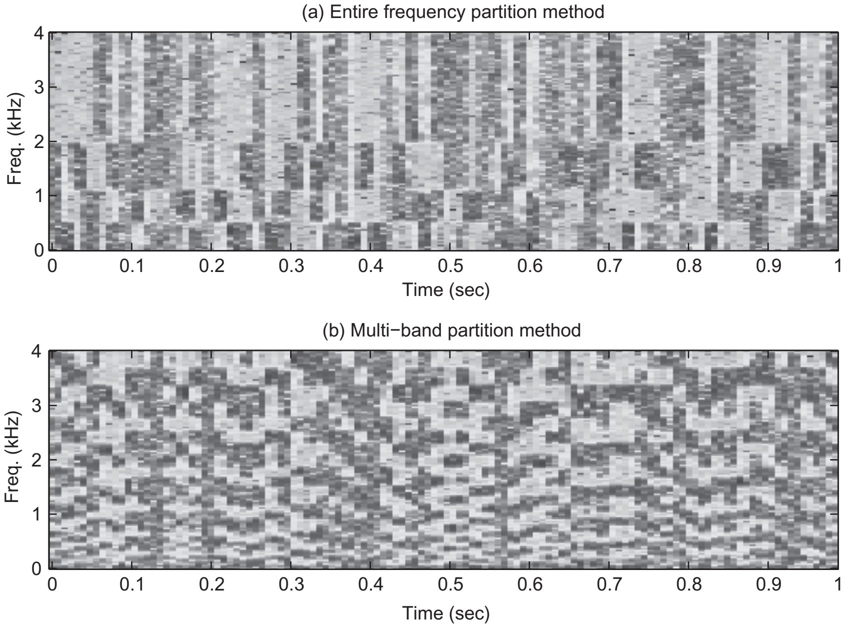

Fig. 3. Spectrograms of colored noise generated by (a) the single-frequency method (i.e. $N = 1$), and (b) the multi-band partition method, with four partitions in each of 23 Mel-frequency bands (i.e. $N = 4$).

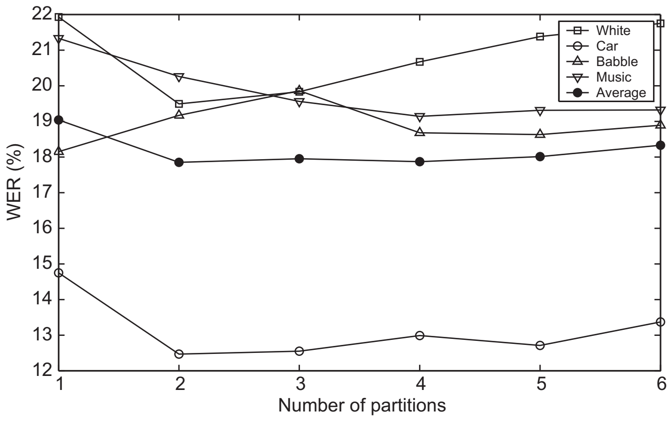

Fig. 4. Recognition performance in average WER over five SNRs (0, 5, 10, 15, and 20 dB) as a function of the number of partitions $N$ used to generate the multi-band colored noise that is used to train the multi-band (i.e. frequency-dependent) mask classifier.

Fig. 4, recognition performance depends somewhat on masker type (with speech babble in the background being a bit of an outlier),[3] and the best overall average performance is observed when the number of frequency partitions is 2 or 4. These trends may be partially accounted for by the nature of the spectral patterns of each noise. As the number of partitions is increased from 1, we believe that performance improves because the mask classification system is exposed to a wider variety of masker spectra in the training process. As the number of partitions continues to increase, however, performance degrades because eventually there remains only a small number of training samples for each masker condition. While the best average WER was obtained with two frequency partitions as 17.85%, it is worth to note that the maskers with four partitions provides better recognition performance for the difficult non-stationary speech babble and music maskers with a comparable average WER as 17.87%.

## 6. Adaptive estimation of *a priori* probabilities for the mask classifier

As described in Section 4.2, the mask classifier presented in this paper consists of a prior model and a GMM formulating a Bayesian classifier. The prior model consists of 23 values between 0 and 1 which represent the prior probabilities of occurrence of the reliable components for each of the Mel-filter-banks. (1.0 minus these values represents the prior probabilities for the corresponding unreliable components.) In the preceding experiments in Sections 4 and 5, these probabilities were estimated by training over the same training data which is used for obtaining the GMM for the classifier. As a result, the prior probabilities that were obtained depend primarily on the acoustic characteristics of the training database. From the results of pilot experiments, we found that matching prior probabilities more closely to the test conditions will produce better speech recognition accuracy. In this section we describe an adaptive method for estimating the prior probabilities.

### 6.1. Estimation of the probabilities of the mask

The distribution of the clean speech feature $X$ in the log-spectral domain is represented by a Gaussian mixture

[3] Recognition accuracy in speech babble is still best with a single partition (i.e. white noise), while the other noise types present the better performance with multiple partitions (i.e. ⩾2) for the colored-noise generation.

model consisting of $K$ components as $(\omega_k, \boldsymbol{\mu}_{X,k}, \boldsymbol{\Sigma}_{X,k})$. A noise model is estimated from non-speech segments in the input speech as a single Gaussian pdf $(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$. In our study, we assume that non-speech segments exist with durations of at least 150 ms at the beginning and end of each input utterance. In the proposed method, we estimate the prior probabilities by assessing the degree of reliability of the mean values of the clean speech model in the log-spectral domain. Here, a reliability estimate $R_k(m)$ of the $m$th frequency band of the $k$th Gaussian component is obtained by comparing a difference in log-spectral value of the mean parameters of clean speech $\mu_{X,k}(m)$ and those of the obtained noise model $\mu_N(m)$ to a threshold as follows:

$$R_k(m) = \begin{cases} 1\,(\text{reliable}), & \text{if } \mu_{X,k}(m) - \mu_N(m) > \zeta, \\ 0\,(\text{unreliable}), & \text{otherwise}, \end{cases} \quad 1 \leqslant m \leqslant M. \tag{6}$$

The parameter $\zeta$ is a threshold for assessing the degree of reliability, and a value of $-2.5$ provided best performance over the noise conditions surveyed in this paper. Finally, the prior probability $Pr(m)$ is obtained by averaging the reliability estimate $R_k(m)$ over all $K$ mixture components:

$$Pr(m) = \frac{1}{K} \sum_{k=1}^{K} R_k(m). \tag{7}$$

In this study we employ separate estimation of the *a priori* probabilities for voiced and unvoiced speech. Using the proposed estimation method, the prior probability for reliable/unreliable components in the test condition can be adaptively estimated as the input speech utterances evolve.

### 6.2. Performance evaluation of the adaptive estimation method of prior probability for mask classifier

In the same fashion as Fig. 4, Fig. 5 describes the recognition accuracy of the frequency-dependent classifier employing the proposed method of estimating the prior probabilities for the masks, as a function of the number of partitions for the colored noise. We note a similar dependence of WER on partition number, with considerably lower WERs observed compared to Fig. 4 over all noise conditions. We obtained the best average WER over all conditions (16.27%) using four partitions, which is an improvement by 1.60% absolute compared to the case when a fixed prior model was used, as in Fig. 4.

Fig. 6 compares directly the impact on recognition accuracy of all of the techniques discussed in this paper. Specifically, we compare the missing-feature reconstruction method employing mask classification with adaptive estimation of prior probabilities (as discussed in this section) to the WER obtained using a baseline classifier (Section 2, Seltzer et al.) features compensation using vector Taylor series (VTS; Moreno et al., 1996) feature compensation, the frequency-dependent classification described in Section 4, the frequency-dependent classification with multi-band artificial noise training (using five bands) described in Section 5, with spectral subtraction (Boll, 1979; Martin, 1994) employed in all approaches. Results are depicted as a function of SNR for the four types of background noise conditions considered in this paper. In all cases the results were obtained without any specific *a priori* knowledge of which masker was used for any particular sentence. These data are also tabulated in Table 2 broken out according to types of background noise and in Table 3 as a function of SNR.

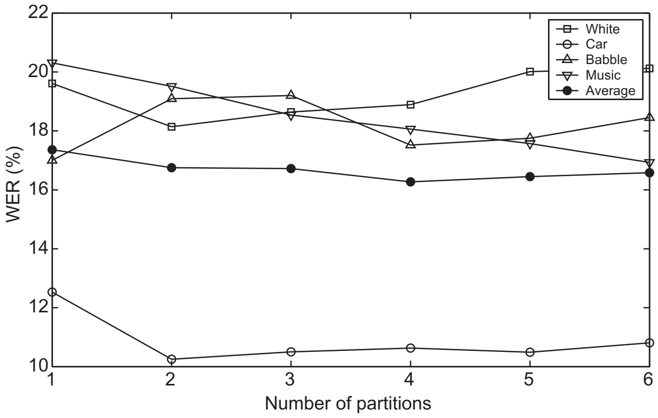

Fig. 5. Word error rates obtained using adaptive estimation of prior mask probabilities averaged over five SNRs (0, 5, 10, 15, and 20 dB) as a function of the number of partitions $N$ used to generate the multi-band colored noise that is used to train the multi-band mask classifier.

The results in Fig. 6 and Tables 2 and 3 demonstrate that the frequency-dependent classifier described in this paper significantly outperforms the baseline classifier except for the white noise condition. (As discussed previously, we expect that the baseline classifier would perform better in white noise because its feature vector exploits more knowledge over the full frequency rage of spectral information and because its model parameters are obtained by training in white noise.) More specifically, the frequency-dependent classifier with multi-band training (FD NM4 + SS) produced an average relative improvement[4] of 9.27% WER over baseline, across all SNRs and all noise conditions (including white noise). The addition of adaptive estimation of the prior probabilities of the masker (AP + FD NM4 + SS) provides an similarly-averaged relative improvement of 25.45% over baseline conditions. These results confirm that the mask classification scheme described in this paper is effective at reducing WER in various types of background noise conditions, when it is employed for missing-feature reconstruction without any prior knowledge of background noise type.

We evaluated the statistical significance of the results described in using the “Matched Pairs Sentence-Segment Word Error (MAPSSWE) Test” provided by NIST.[5] Except for the case of comparisons to baseline performance

[4] The average relative improvement is computed by taking the average of the obtained relative improvements.

[5] <http://www.nist.gov/speech/tests/sigtests/mapsswe.htm>.

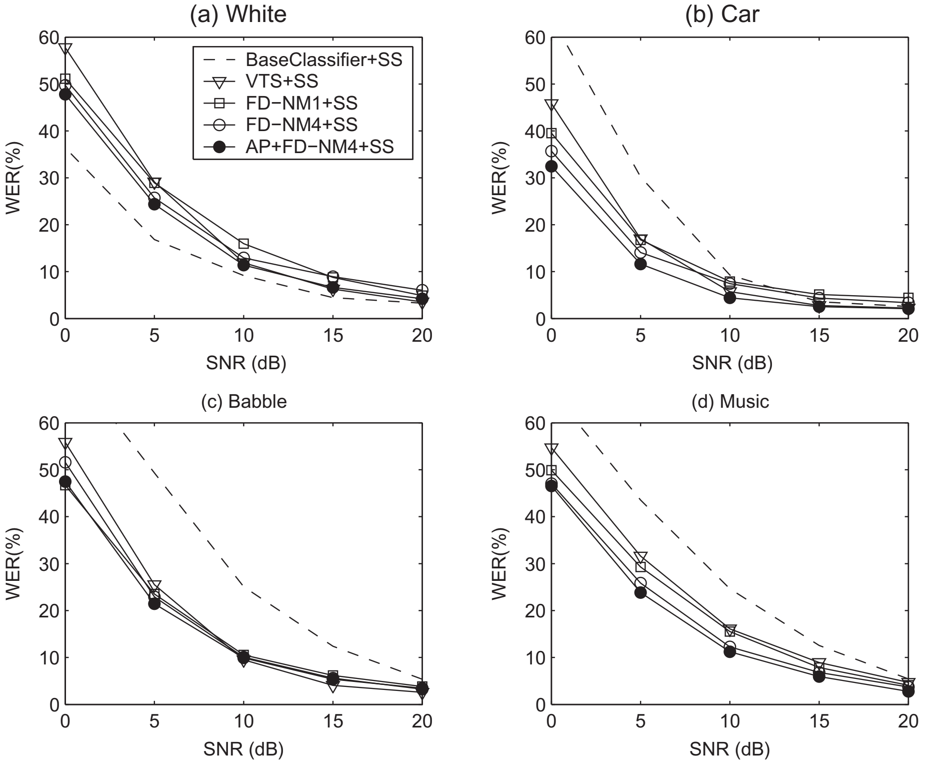

Fig. 6. Word error rates obtained using our missing-feature approach employing a baseline classifier including spectral subtraction (SS), vector Taylor series (VTS), frequency-dependent mask classification trained on white noise (FD NM1), multi-band artificial colored noise (FD NM4), and adaptive estimation of the mask probabilities (AP FD NM4), all as a function of SNR.

Table 2
Comparison of WER (%) in four types of background noises averaged over all SNRs (0, 5, 10, 15, and 20 dB).

| | White | Car | Babble | Music | Average |
|---|---|---|---|---|---|
| No processing | 34.26 | 38.20 | 46.60 | 34.15 | 38.30 |
| SS | 20.88 | 25.39 | 43.22 | 33.50 | 30.75 |
| VTS + SS | 21.72 | 14.71 | 19.49 | 23.19 | 19.78 |
| Baseline classifier + SS | 14.03 | 21.97 | 33.21 | 30.60 | 24.95 |
| FD NM4 + SS | 20.67 | 12.99 | 18.68 | 19.14 | 17.87 |
| (Average relative improvement) | (−62.90) | (+12.85) | (+47.51) | (+39.61) | (+9.27) |
| AP + FD NM4 + SS | 18.89 | 10.63 | 17.52 | 18.06 | 16.27 |
| (Average relative improvement) | (−35.68) | (+42.11) | (+49.04) | (+46.35) | (+25.45) |

Table 3
Comparison of WER (%) as a function of SNR conditions averaged over the four types of background noise.

| | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB | Average |
|---|---|---|---|---|---|---|
| No processing | 82.79 | 58.74 | 31.46 | 13.33 | 5.19 | 38.30 |
| SS | 69.76 | 43.96 | 23.19 | 11.36 | 5.48 | 30.75 |
| VTS + SS | 53.52 | 25.81 | 10.80 | 5.50 | 3.26 | 19.78 |
| Baseline classifier + SS | 60.33 | 34.98 | 17.04 | 8.25 | 4.16 | 24.95 |
| FD NM4 + SS | 46.01 | 22.14 | 10.68 | 6.42 | 4.10 | 17.87 |
| (Average relative improvement) | (+16.89) | (+23.74) | (+22.12) | (−5.12) | (−11.29) | (+9.27) |
| AP + FD NM4 + SS | 43.57 | 20.31 | 9.24 | 5.11 | 3.14 | 16.27 |
| (Average relative improvement) | (+21.05) | (+29.68) | (+35.70) | (+22.49) | (+18.35) | (+25.45) |

in the presence of white noise, virtually all of the techniques discussed in Sections 4–6 were effective at a level of significance of $p = .001$ at SNRs of 0 and 5 dB, and at a level of $p = .05$ at the higher SNRs considered.

## 7. Conclusions

In this paper we have described several useful improvements to the process of estimating masks for speech recognition systems that employ missing-feature restoration. The innovations we describe provide better speech recognition accuracy for systems operating in the presence of unknown background noise. We first described an alternate approach to the Bayesian classification of missing-feature masks in which the masks in each spectral subband were developed independently of other subbands. The proposed mask classifier employs a set of frequency-dependent features including subband cepstral coefficients, spectral flatness measure, and comb filter ratio. We also proposed a new method of generating colored-noise signals to train the frequency-dependent mask classifier, which is based on masking noises with artificially-introduced spectral and temporal variations. These signals simulate a number of spectral patterns with a relatively small amount of training data within the proposed mask classifier framework. To obtain a characterization of the prior probabilities for the mask classifier, we also proposed an adaptive estimation method. Our mask classification scheme was evaluated in the context of a speech recognition system that exploits missing-feature reconstruction. The experimental results showed that the frequency-dependent mask classification trained using the artificial colored-noise signals, and employing the adaptive estimation of the probabilities of the mask classifier, is effective in improving speech recognition accuracy in the presence of various types of noise maskers, especially at lower SNRs, without any prior knowledge of the test conditions.

## Acknowledgments

This work was supported by the Postdoctoral Fellow Program of the Korea Science and Engineering Foundation (KOSEF) and by the National Science Foundation (Grant IIS-0420866). Much of the work described here was previously presented at the Interspeech 2005 (Kim et al., 2005) and IEEE ICASSP 2006 (Kim and Stern, 2006) conferences, although with somewhat different notation.

## References

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: ICSPL-2000, pp. 373–376.

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

Cooke, M., Morris, A., Green, P., 1997. Missing data techniques for robust speech recognition. In: ICASSP-97, pp. 863–866.

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34 (3), 267–285.

Davis, S.B., Mermelstein, P., 1980. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. 28 (4), 357–366.

ETSI standard document, 2000. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithms. ETSI ES 201 108 v1.1.2 (2000–04).

Harding, S., Barker, J., Brown, G.J., 2005. Mask estimation based on sound localisation for missing data speech recognition. In: ICASSP-2005, pp. 537–540.

Hirsch, H.G., Pearce, D., 2000. The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions. In: ISCA ITRW ASR-2000.

Jancovic, P., Kokuer, M., Murtagh, F., 2003. High-likelihood model based on reliability statistics for robust combination of features: application to noisy speech recognition. In: Eurospeech-2003, pp. 2161–2164.

Johnston, J.D., 1988. Transform coding of audio signals using perceptual noise criteria. IEEE J. Selected Areas Commun. 6 (2), 314–323.

Josifovski, L., Cooke, M., Green, P., Vizihno, A., 1999. State based imputation of missing data for robust speech recognition and speech enhancement. In: Eurospeech-99, pp. 2837–2840.

Kim, W., Stern, R.M., Ko, H., 2005. Environment-independent mask estimation for missing-feature reconstruction. In: Interspeech-2005, pp. 2637–2640.

Kim, W., Stern, R.M., 2006. Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise. In: ICASSP-2006, pp. 305–308.

Lippmann, R.P., Carlson, B.A., 1997. Using missing-feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise. In: Eurospeech-97, pp. 37–40.

Martin, R., 1994. Spectral subtraction based on minimum statistics. In: EUSIPCO-94, pp. 1182–1185.

Moreno, P.J., Raj, B., Stern, R.M., 1996. A vector Taylor series approach for environment-independent speech recognition. In: ICASSP-96, pp. 733–736.

Park, H.-M., Stern, R.M., 2009. Spatial separation of speech signals using amplitude estimation based on interaural comparisons of zero crossings. Speech Commun. 51 (1), 15–25.

Raj, B., Seltzer, M.L., Stern, R.M., 2004. Reconstruction of missing features for robust speech recognition. Speech Commun. 43 (4), 275–296.

Raj, B., Singh, R., 2005. Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition. In: ASRU-2005, pp. 65–70.

Raj, B., Stern, R.M., 2005. Missing-feature approaches in speech recognition. IEEE Signal Process. Magazine 22 (5), 101–116.

Renevey, P., Drygajlo, A., 2001. Detection of reliable features for speech recognition in noisy conditions using a statistical criterion. In: CRAC2001.

Seltzer, M.L., 2000. Automatic Detection of Corrupted Speech Features for Robust Speech Recognition. M.S. Thesis, Carnegie Mellon University.

Seltzer, M.L., Raj, B., Stern, R.M., 2004. A Bayesian classifier for spectrographic mask estimation for missing-feature speech recognition. Speech Commun. 43 (4), 379–393.

Singh, R., Stern, R.M., Raj, B., 2002a. Model compensation and matched condition methods for robust speech recognition. Chapter in CRC Handbook on Noise Reduction in Speech Applications. CRC Press.

Singh, R., Stern, R.M., Raj, B., 2002b. Signal and feature compensation methods for robust speech recognition. Chapter in CRC Handbook on Noise Reduction in Speech Applications. CRC Press.

Srinivasan, S., Roman, N., Wang, DeL., 2006. Binary and ratio time-frequency masks for robust speech recognition. Speech Commun. 48 (11), 1486–1501.