Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 48 (2006) 1486–1501

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Binary and ratio time-frequency masks for robust speech recognition

Soundararajan Srinivasan [a,*], Nicoleta Roman [b], DeLiang Wang [b]

[a] *Biomedical Engineering Department, The Ohio State University, 395 Dreese Laboratories, 2015 Neil Avenue, Columbus, OH 43210, USA*
[b] *Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210, USA*

Received 14 June 2005; received in revised form 11 September 2006; accepted 14 September 2006

Abstract

A time-varying Wiener filter specifies the ratio of a target signal and a noisy mixture in a local time-frequency unit. We estimate this ratio using a binaural processor and derive a ratio time-frequency mask. This mask is used to extract the speech signal, which is then fed to a conventional speech recognizer operating in the cepstral domain. We compare the performance of this system with a missing-data recognizer that operates in the spectral domain using the time-frequency units that are dominated by speech. To apply the missing-data recognizer, the same binaural processor is used to estimate an ideal binary time-frequency mask, which selects a local time-frequency unit if the speech signal within the unit is stronger than the interference. We find that the performance of the missing data recognizer is better on a small vocabulary recognition task but the performance of the conventional recognizer is substantially better when the vocabulary size is increased.

Keywords: Ideal binary mask; Ratio mask; Robust speech recognition; Missing-data recognizer; Binaural processing; Speech segregation

1. Introduction

The performance of automatic speech recognizers (ASRs) degrades rapidly in the presence of noise, microphone variations and room reverberation (Gong, 1995; Lippmann, 1997). Speech recognizers are typically trained on clean speech and face a problem of mismatch when used in conditions where speech occurs simultaneously with other sound sources. To mitigate the effect of this mismatch on recognition, noisy speech is typically preprocessed by speech enhancement algorithms, such as microphone arrays (Bradstein and Ward, 2001; Cardoso, 1998; Ehlers and Schuster, 1997; Hughes et al., 1999), computational auditory scene analysis (CASA) systems (Brown and Wang, 2005; Rosenthal and Okuno, 1998) or spectral subtraction techniques (Boll, 1979; Droppo et al., 2002). Microphone arrays require the number of sensors to increase as the number of interfering sources increases. Monaural CASA systems employ harmonicity as the primary

[*] Corresponding author. Tel.: +1 614 292 7402.
E-mail addresses: srinivasan.36@osu.edu (S. Srinivasan), roman.45@osu.edu (N. Roman), dwang@cse.ohio-state.edu (D.L. Wang).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.09.003

cue for grouping acoustic components corresponding to speech. These systems, however, do not perform in time-frequency (T-F) regions that are dominated by unvoiced speech. Spectral subtraction systems typically assume stationary noise. Hence, in the presence of non-stationary noise sources, their performance is not adequate for recognition (Cooke et al., 2001). If samples of the corrupting noise source are available *a priori,* a model for the noise source can additionally be trained and noisy speech may be jointly decoded using the trained models of speech and noise (Gales and Young, 1996; Varga and Moore, 1990) or enhanced using linear filtering methods (Ephraim, 1992). However, in many realistic applications, adequate amounts of noise samples are not available *a priori* and hence training of a noise model is not feasible.

Recently, a missing-data approach to speech recognition in noisy environments has been proposed by Cooke et al. (2001). This method is based on distinguishing between reliable and unreliable data. When speech is contaminated by additive noise, some time-frequency units contain predominantly speech energy (reliable) and the rest are dominated by noise energy. The missing-data method treats the latter T-F units as missing or unreliable during recognition (see Section 4.2). Missing T-F units are identified by thresholding the T-F units based on local SNR. Spectral subtraction is typically used to estimate the local SNR. The performance of the missing-data recognizer is significantly better than the performance of a system using spectral subtraction for speech enhancement followed by recognition of enhanced speech (Cooke et al., 2001).

A potential disadvantage of the missing-data recognizer is that recognition is performed in the spectral or T-F domain. It is well known that recognition using cepstral coefficients yields a superior performance compared to recognition using spectral coefficients under clean speech conditions (Davis and Mermelstein, 1980). The superiority of the cepstral features stems from the ability of the cepstral transformation to separate vocal-tract filtering from excitation source in speech production (Rabiner and Juang, 1993). Additionally, the cepstral transform approximately orthogonalizes the spectral features (Shire, 2000). Since the missing-data recognition is based on marginalizing the unreliable T-F features during recognition, it is coupled with a spectral or T-F representation. Any global transformation of the spectral features (e.g. cepstral transformation) smears the information from the noisy T-F units across all the global features, preventing its effective marginalization. Attempts to adapt the missing-data method to the cepstral domain have centered around reconstruction or imputation of the missing values in the spectral domain followed by transformation to the cepstral domain (Cooke et al., 2001; Raj et al., 2004). Alternatively, van Hamme (2003) performs imputation directly in the cepstral domain. These reconstructions are typically based either on the speech recognizer itself or on other trained models of speech. The success of these model-based imputation techniques depend on the adequacy of reliable data for identification of the correct speech model for imputation. In addition, errors in imputation procedures affect the performance of the system even when the model is correctly identified.

Another potential drawback of the missing-data recognizer, which has not been well studied, is the problem of data paucity. The amount of "reliable" data available to the recognizer is a function of both SNR and frequency characteristics of the noise source. A decrease in SNR, as well as an increase in the bandwidth of the noise source causes an increase in the amount of missing data. This leads to a deterioration in performance for a small vocabulary task (Cooke et al., 2001). The reduction in reliable data may pose an additional problem for recognition with larger vocabulary sizes. Paucity of reliable data constrains the missing-data recognizer to use only a small portion of the total T-F acoustic model space. This reduced space may be insufficient to differentiate between a large number of competing hypotheses during decoding. In this paper, we study this issue by comparing the performance of the missing-data recognizer on two tasks with different vocabulary sizes.

Binaural CASA systems that compute binary masks have been used successfully as front-ends for the missing-data recognizer on small vocabulary tasks (Palomaki et al., 2004; Roman et al., 2003). Such systems compare the acoustic signals at the two ears in order to extract the binaural cues of interaural time differences (ITD) and interaural intensity differences (IID). These binaural cues are correlated with the location of a sound source and hence provide powerful mechanisms for segregating sound sources from different locations. Moreover, binaural processing is independent of the signal content and hence can be used to segregate both voiced and unvoiced speech components from a noisy mixture. The computational goal of the binaural

CASA systems is an ideal binary mask. A T-F unit in the ideal binary mask is labeled 1 or reliable if the corresponding T-F unit of the noisy speech contains more speech energy than interference energy; it is labeled 0 or unreliable otherwise.[1] We employ a recent binaural speech segregation system (Roman et al., 2003) to estimate an ideal binary T-F mask. This mask is fed to the missing-data recognizer and recognition is performed in the spectral domain.

The minimum mean-square error (MMSE) based short-time spectral amplitude estimator, which utilizes *a priori* SNR in a local T-F unit, has been used previously to effectively enhance noisy speech (Ephraim and Malah, 1984). *A priori* SNR can be obtained if premixing speech and noise signals are available. Roman et al. (2003) have shown that in a narrow frequency band, there exists a systematic relationship between *a priori* SNR and values of ITD and IID. Motivated by this observation, we estimate an ideal ratio T-F mask using statistics collected for ITD and IID at each individual frequency bin. A unit in the ratio mask is a measure of the speech energy to total energy (speech and noise) in the corresponding T-F unit of the noisy signal. The ratio mask is then used to enhance the speech, enabling recognition using Mel-frequency cepstral coefficients (MFCCs). We use "conventional recognizer" to refer to a continuous density hidden Markov model (HMM) based ASR using MFCCs as features.

We compare the performance of the conventional recognizer to that of the missing-data recognizer on a robust speech recognition task. In particular, we examine the effect of vocabulary size on the performance of the two recognizers. We find that on a small vocabulary task, the missing-data recognizer outperforms the conventional ASR. Our finding is consistent with a previous comparison using a binaural front-end made on a small vocabulary "cocktail-party" recognition task (Glotin et al., 1999; Tessier et al., 1999). The accuracy of results obtained using the missing-data method in the spectral domain was reported to be better than those obtained using the conventional ASR in the cepstral domain. With an increase in the vocabulary size, however, the conventional ASR performs substantially better. Results using the missing value imputation methods have been reported on a larger vocabulary previously (Raj et al., 2004). Their method uses a binary mask and therefore is subject to the same limitations stated previously.

The rest of the paper is organized as follows. Section 2 provides an overview of the proposed systems. We then describe the binaural front-end for both the conventional and missing-data recognizers in Section 3. The section additionally provides the estimation details of ideal binary and ratio T-F masks. The conventional and missing-data recognition methods are reviewed in Section 4. The recognizers are tested on two different task domains with different vocabulary sizes. Section 5 discusses the two tasks and presents the evaluation results of the recognizers along with a comparison of their relative performance. Finally, conclusion and future work are given in Section 6.

2. System overview

In this study, we analyze two strategies for robust speech recognition: (1) missing-data recognition and (2) a system that combines speech enhancement with a conventional ASR. The performance is examined at various SNR conditions and for two vocabulary sizes. Fig. 1 shows the architecture of the two different processing strategies.

The input to both systems is a binaural mixture of speech and interference presented at different, but fixed, locations. The measurements of head-related transfer functions (HRTFs) are a standard method for binaural synthesis. HRTF encompasses the filtering effects of head, torso and pinna, and encodes information corresponding to the source location. In this paper, the HRTFs are provided by the measured left and right responses of the KEMAR manikin from a distance of 1.4 m in the horizontal plane, resulting in two 128 point impulse responses at a sampling rate of 44.1 kHz (Gardner and Martin, 1994). We generate the left and right ear signals by filtering the monaural signals with the left and right HRTFs. Note that different locations provide different HRTFs, resulting in different binaural signals. The responses to multiple sources are added at each ear. Due to the differential filtering effects between the two ears, the HRTFs provide location-dependent ITD and IID which can be extracted independently in each T-F unit (see Section 3). The T-F resolution is 20 ms time frames

[1] Similar masks have also been referred to as '*a priori* masks' and 'oracle masks' in the literature.

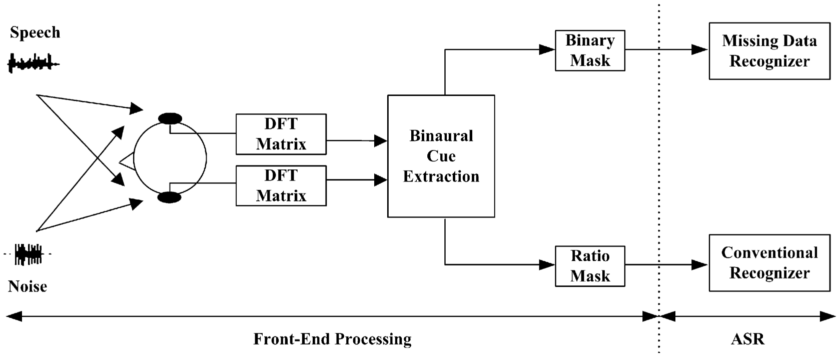

Fig. 1. Architecture of the two robust speech recognition strategies with binaural preprocessing: The missing-data recognizer and the conventional ASR. Left and right ear signals are obtained by filtering with HRTFs. A short-time Fourier analysis is applied to the signals, resulting in a time-frequency decomposition. ITD and IID are computed in each T-F unit. The missing-data recognizer works with a binary mask. A ratio mask is used as a speech enhancement strategy and is fed to the conventional recognizer.

with a 10 ms frame shift, and 512 DFT coefficients. Frames are extracted by applying a running Hamming window to the signal. The frequency bins uniformly cover the complete frequency range, up to the Nyquist frequency.

The missing-data speech recognizer operates in the log-spectral domain and requires a binary mask that informs the recognizer of which T-F units of the noisy input are dominated by speech energy. A 64-channel auditory filterbank was used previously as a front-end for the missing-data recognizer (Cooke et al., 2001). We have chosen a DFT representation for the missing-data recognizer in order to be consistent with the conventional recognizer. A comparison between the DFT representation and the auditory filterbank representation has shown that the difference in recognition performance is statistically insignificant. Statistics based on mixtures of multiple speech sources show that there exists a systematic correlation between the *a priori* energy ratio and the estimated ITD/IID values, resulting in a characteristic clustering across frequencies (Roman et al., 2003). To estimate the ideal binary mask we extend the non-parametric classification method of Roman et al. (2003) in the joint ITD/IID feature space independently for each frequency bin. The frequency decomposition used by Roman et al. (2003) was generated by a gammatone filterbank. This classification results in binary Bayesian decision rules that determine whether speech is stronger than interference in individual T-F units (energy ratio greater than 0.5). The system of Roman et al. (2003) was chosen because of the excellent match between their estimated binary mask and the ideal binary mask.

The conventional approach to robust speech recognition involves preprocessing of the corrupted speech by speech enhancement algorithms. This allows for the subsequent usage of decorrelating transformations (cepstral transformation, linear discriminant analysis), temporal processing methods (delta features, RASTA filtering) and normalization techniques (mean subtraction, variance normalization) on enhanced spectral features (Chen et al., 2005; Shire, 2000). In this study we use cepstral and delta features along with cepstral mean subtraction, which are known to provide improved recognition accuracy. To enhance noisy speech we estimate a ratio T-F mask. The statistics described above show that the estimated ITD and IID have a functional relationship with the *a priori* energy ratio. We employ this relationship in a non-parametric fashion to estimate the ideal ratio mask. Finally, to decode using the conventional ASR, MFCCs are computed from speech enhanced by masking the corrupted signal with the estimated ratio mask.

We evaluate the performance of the missing-data recognizer using both ideal and estimated binary masks. This is then compared to that of the conventional ASR using both ideal and estimated ratio masks.

3. A localization based front-end for ASR

When speech and additive noise are orthogonal, the linear MMSE filter is the Wiener filter (van Trees, 1968). With a frame-based processing, the MMSE filter corresponds to the ratio of speech eigenvalues to the sum of eigenvalues of speech and noise (van Trees, 1968). The eigenvalues can be computed from the auto-covariance functions prior to mixing by considering speech and noise to be two distinct random processes. Under asymptotic conditions, the MMSE filter corresponds to the frame-based Wiener filter (McAulay and Malpass, 1980; van Trees, 1968). Ephraim and Malah (1984) have additionally shown that the optimal MMSE estimate of speech spectral amplitude in a local T-F unit is strongly related to the *a priori* SNR. To estimate the speech in a local T-F unit, we approximate the frame-based filter with an ideal ratio mask defined using the *a priori* energy ratio $R(\omega,t)$:

$$R(\omega,t)=\left[\frac{|S(\omega,t)|^2}{|S(\omega,t)|^2+|N(\omega,t)|^2}\right], \tag{1}$$

where $S(\omega,t)$ and $N(\omega,t)$ are the target and noise spectral values at frequency ω and time t computed from the signal at the "better ear" – the ear with higher SNR. Our computational goal for front-end processing with the conventional ASR is to estimate $R(\omega,t)$ directly from the noisy mixture.

In addition, we define the ideal binary T-F mask as:

$$B(\omega,t)=\begin{cases}1 & \text{if } R(\omega,t)\geqslant\theta,\\ 0 & \text{otherwise},\end{cases} \tag{2}$$

where θ is set to be 0.5. Such masks have been shown to generate high-quality reconstruction for a variety of signals and also provide an effective front-end for missing-data recognition on a small vocabulary task (Cooke et al., 2001; Roman et al., 2003).

The objective of our front-end processing is to develop effective mechanisms for estimating both ideal binary and ratio masks. We propose an estimation method based on observed patterns for the binaural cues caused by the auditory interaction of multiple sources presented at different locations. Roman et al. (2003) have shown that ITD and IID undergo systematic shifts as the energy ratio between the target and the interference changes. In a particular frequency bin, the ITD and IID corresponding to the target source exhibit location dependent characteristic values. As the SNR in this frequency bin decreases due to the presence of interference, the ITD and IID systematically shift away from the target values. Theoretical derivations for the case of two sinusoidal sources can be found in Roman et al. (2003). Moreover, statistics collected from real signals have shown similar patterns. In particular, the empirical mean approximates well the theoretical mean obtained in the case of two sinusoids (Roman et al., 2003). Note that training for each frequency bin is required since frequency-dependent combinations of ITD and IID arise naturally for a fixed spatial configuration. We employ the same training corpus as used by Roman et al. (2003) consisting of 10 speech signals from the TIMIT database (Garofolo et al., 1993). Five sentences correspond to the target location set and the rest belong to the interference location set. Binaural signals are obtained by convolving with KEMAR HRTFs as described in Section 2. This dataset is different from the databases used in training the ASRs.

The ITD/IID estimates are computed independently in each T-F unit based on the spectral ratio at the left and right ears:

$$(I\hat{T}D,I\hat{I}D)(\omega,t)=\left[-\frac{1}{\omega}A\left(\frac{X_{\mathrm{L}}(\omega,t)}{X_{\mathrm{R}}(\omega,t)}\right),\frac{|X_{\mathrm{L}}(\omega,t)|}{|X_{\mathrm{R}}(\omega,t)|}\right], \tag{3}$$

where $X_{\mathrm{L}}(\omega,t)$ and $X_{\mathrm{R}}(\omega,t)$ are the left and right ear spectral values of the noisy speech at frequency ω and time t and $A(re^{\mathrm{j}\phi})=\phi$, $-\pi<\phi\leqslant\pi$. The function $A(\cdot)$ computes the phase angle, in radians, of a complex number with magnitude r and phase angle ϕ. Note that the phase is ambiguous corresponding to integer multiples of 2π. To disambiguate, we identify ITD in the range of $2\pi/\omega$ centered at zero delay. By dividing the relative phase angle by the radian frequency, $I\hat{T}D$ estimates the time difference between the left and the right ear signals. $I\hat{I}D$ calculates the relative magnitude between the left and the right ear spectral values and hence estimates the intensity difference.

3.1. Estimation of the binary mask

Fig. 2 shows empirical results from the training corpus for a two-source configuration: target source in the median plane and interference at 30°. In the training corpus, we have access to target and interference signals separately. Hence, we can compute R for each mixture signal. The scatter plot in Fig. 2(a) shows samples of $I\hat{T}D$ and R, as well as the mean, the standard deviation, and the histogram (the bottom panel), for a frequency bin at 1 kHz. Similarly, Fig. 2(b) shows the results that describe the variation of $I\hat{I}D$ and R for a frequency bin at 3.4 kHz. The results are similar to those obtained by Roman

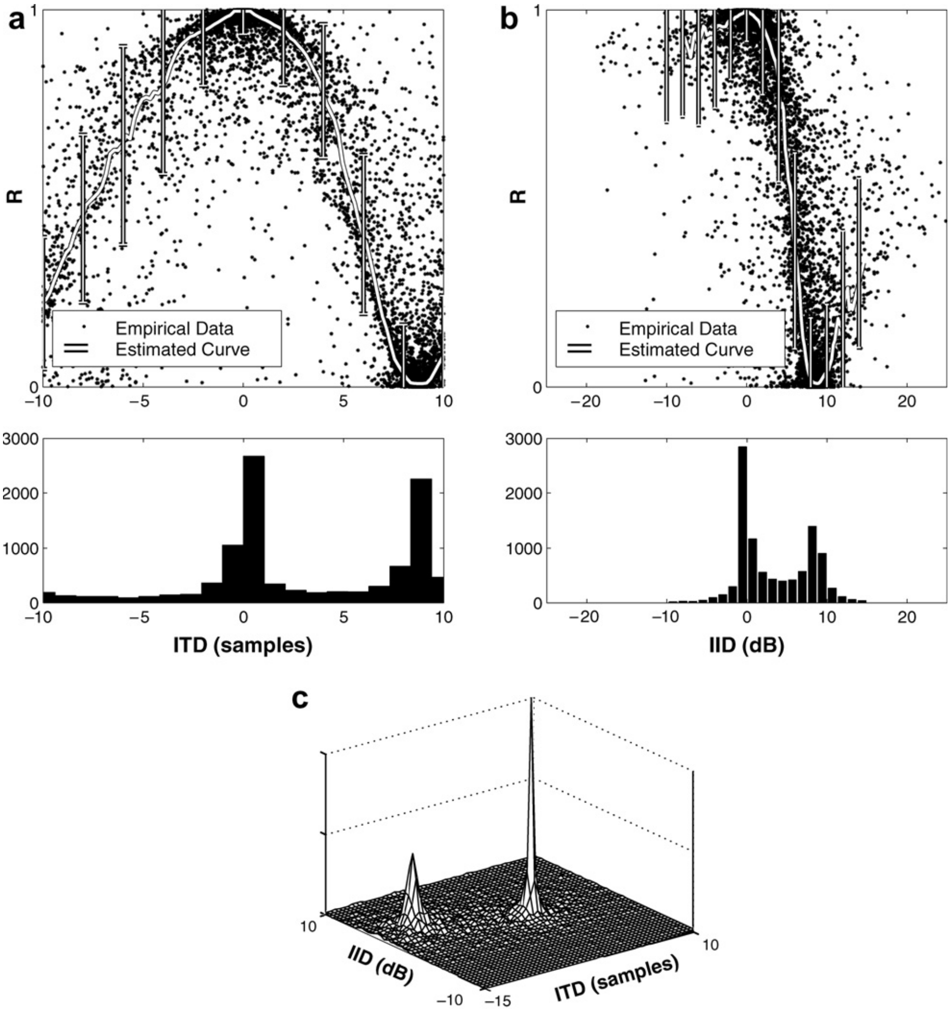

Fig. 2. Relationship between ITD/IID and the energy ratio R. Statistics are obtained with target in the median plane and interference on the right side at 30°. (a) The top panel shows the scatter plot for the distribution of R with respect to ITD for a frequency bin at 1 kHz. The solid white curve shows the mean curve fitted to the data. The vertical bars represent the standard deviation. The bottom panel shows the histogram of ITD samples. (b) Corresponding results for IID for a frequency bin at 3.4 kHz. (c) Histogram of ITD and IID samples for a frequency bin at 2 kHz.

et al. (2003), who use an auditory filterbank for frequency decomposition. Note that for the T-F units dominated by target ($R \approx 1$), the binaural cues are clustered around the target values. Similarly, for the T-F units dominated by interference ($R \approx 0$), the binaural cues are clustered around the interference values. Furthermore, the scatter plots exhibit a systematic shift of the estimated ITD and IID as R, as defined in (1), varies from 1 to 0. Moreover, a location-based clustering is observed in the joint ITD–IID space as shown in Fig. 2(c). Each peak in the histogram corresponds to a distinct active source. Therefore, to estimate the binary mask $\hat{B}(\omega,t)$, we employ non-parametric classification in the joint ITD–IID feature space as used by Roman et al. (2003). There are two hypotheses for the binary decision: H_1 – target is stronger or $R \geqslant 0.5$ and H_2 – interference is stronger or $R < 0.5$. The estimated binary mask, $\hat{B}(\omega,t)$, is obtained using the maximum a posteriori (MAP) decision rule:

$$\hat{B}(\omega,t) = \begin{cases} 1 & \text{if } p(H_1)p(x|H_1) > p(H_2)p(x|H_2), \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

where x is the $(I\hat{T}D, I\hat{I}D)(\omega,t)$ feature vector. The prior probabilities, $p(H_i)$, are computed as the ratio of the number of samples in each class to the total number of samples. The conditional probabilities, $p(x|H_i)$, are estimated from the training data using the kernel density estimation method (Roman et al., 2003).

3.2. Estimation of the ratio mask

In order to estimate the ideal ratio mask, we use the same training data. It is well known that ITD is salient at low frequencies while IID becomes more prominent at higher frequencies (Blauert, 1997). In the high-frequency range, the wavelength of the acoustic signal is much shorter compared to the distance between the two ears. This results in a compressed range for the unambiguous values of ITD, which reduces the discriminative power of this cue. On the other hand, while the range of IID is very small at low frequencies, it can be as high as 30 dB at high frequencies. Hence, IID is a more reliable cue at high frequencies.

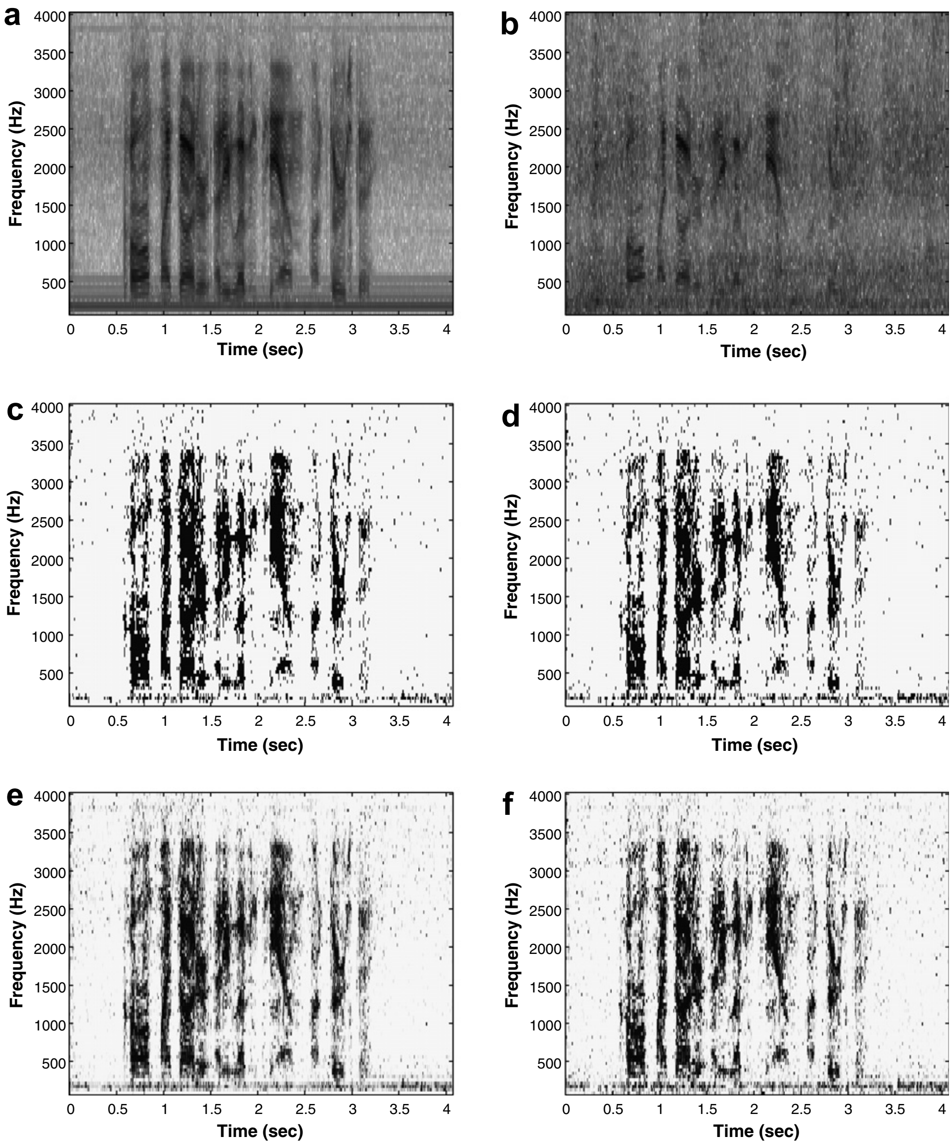

Fig. 3. Comparison between estimated and ideal binary and ratio T-F masks for a mixture of a clean speech utterance presented in the median plane and an interference signal presented at 30°. The SNR is 0 dB. (a) Spectrogram of the clean speech utterance. (b) Spectrogram of the mixture. (c) The ideal binary mask. (d) The estimated binary mask. (e) The ideal ratio mask. (f) The estimated ratio mask. The signals correspond to the left ear.

ITD exhibits different patterns across frequency bins as seen in the number of modes that characterizes the distribution of the samples (Roman et al., 2003). Hence, there is no unique parametric curve for all frequencies. Moreover, in the absence of evidence for a parametric estimate to provide better recognition results, a mean curve is fitted to the distribution of ITD. This is our estimated ratio mask below 3 kHz. For higher frequencies, we utilize the information provided by the IID cues and use the same method to estimate the energy ratio. For improved results, we remove the outliers outside of 0.2 distance from the median. The resulting mean curves are shown in Figs. 2(a) (ITD) and (b) (IID). Thus, for given $I\hat{T}D(\omega,t)$ and $I\hat{I}D(\omega,t)$, the estimated energy ratio $\hat{R}(\omega,t)$ is the corresponding value on the mean curve.

Fig. 3 shows the comparison between ideal and estimated masks. Figs. 3(a) and (b) show the spectrograms of a clean speech utterance and the noisy mixture, respectively. The noisy mixture is generated by combining the clean speech utterance presented in the median plane with a factory noise signal presented on the right side at 30°. The SNR is 0 dB. The mask estimation algorithms described above are applied to the mixture and the results are presented in Fig. 3(c)–(f). Figs. 3(c) and (d) show the ideal binary mask and the estimated binary mask, respectively. Notice that the estimated binary mask approximates well the ideal binary mask (see also Roman et al., 2003). Figs. 3(e) and (f) show the ideal ratio mask and the estimated ratio mask, respectively. Observe that the estimated ratio mask is very similar to the ideal ratio mask, especially in the high speech energy T-F units.

4. Recognition strategies

We evaluate the binaural segregation system described in Section 3 as the front-end for robust ASR using two different recognizers. Conventional ASR uses MFCCs as the parameterization of observed speech. MFCCs are computed from the segregated speech obtained after applying the ratio mask to the noisy input signal (see Eq. (5)). The missing-data recognizer uses log-spectral energy as feature vectors in conjunction with the binary mask, generated by the binaural system. An HMM toolkit, HTK (Young et al., 2000), is used in the training of both recognizers and the testing with the conventional ASR. During testing with the missing-data recognizer, the decoder is modified to incorporate the missing-data methods.

4.1. The conventional speech recognizer

We use the standard continuous density HMM based speech recognizer trained on clean speech to model each word in the vocabulary (Section 5). Observation densities are modeled as a mixture of Gaussians with diagonal covariance. The input to this ASR is the estimated speech spectral energy $|\hat{S}(\omega,t)|^2$, obtained by an element-wise multiplication of the estimated ratio mask and the spectral energy of the noisy speech.

$$|\hat{S}(\omega,t)|^2 = |X(\omega,t)|^2 \cdot \hat{R}(\omega,t), \tag{5}$$

where $|X(\omega,t)|^2$ is the spectral energy of the noisy signal at the better ear (see Section 5). One of our evaluation corpora (see Section 5) is originally band-limited to 4 kHz. Hence, we apply a rectangular window to the estimated spectra and truncate them to 4 kHz. These truncated speech spectra are processed by a Mel-frequency filterbank (Rabiner and Juang, 1993), comprising 26 triangle-shaped filters. The low frequency cut-off of the first filter is set to 105 Hz and the high frequency cut-off of the last filter was set to 4300 Hz. A log compression is then applied to the resulting spectra. Finally, the spectral coefficients are converted to cepstral coefficients via the discrete cosine transform (Oppenheim et al., 1999). The performance of the conventional ASR with full-band spectra is shown in an earlier study (Srinivasan et al., 2004). We observe that the application of the rectangular window eliminates the effect of the spurious and inaccurately estimated high-frequency components. This helps significantly improve accuracy compared to recognition using full-band spectra.

MFCCs are chosen as feature vectors as they are most commonly used in state-of-the-art recognizers (Rabiner and Juang, 1993). Thirteen cepstral coefficients along with delta and acceleration coefficients are extracted each frame, including the 0th order cepstral coefficient C_0 as the energy term. Frames are extracted as described in Section 2. A first-order preemphasis coefficient of 0.97 is applied to the signal. Cepstral mean normalization is additionally implemented for improved ASR performance (Young et al., 2000).

4.2. The missing-data speech recognizer

The missing-data recognizer (Cooke et al., 2001) is an HMM-based ASR that makes use of spectro-temporal redundancy in speech to recognize noisy speech based on its speech dominant T-F units. Given an observed speech vector Y, the problem of word recognition is to maximize the posterior $P(W_i|Y)$, where W_i is a valid word sequence according to the grammar for the recognition task. When parts of Y are corrupted by additive noise, it can be partitioned into its reliable and unreliable constituents as Y_r and Y_u. One can then seek the Bayesian decision given the reliable constituents. In the marginalization method, the posterior probability using only the reliable constituents is computed by integrating over the unreliable ones (Cooke et al., 2001). In missing-data methods, recognition is typically performed using spectral energy as feature vectors. If Y represents the observed spectrum and sound sources are additive and uncorrelated, then the unreliable parts may be constrained as $0 \leqslant \widetilde{Y}_\mathrm{u}^2 \leqslant Y_\mathrm{u}^2$. This constraint, therefore, states that the true value of the spectral energy in the unreliable parts, $\widetilde{Y}_\mathrm{u}^2$, lies between 0 and the observed spectral energy. These bounds are used as limits on the integral involved in marginalizing the posterior probability over the unreliable features. This bounded marginalization method is shown by Cooke et al. (2001) to have a better recognition score than the simple marginalization method, and is hence used in all our experiments employing the missing-data recognizer. We use mixture of Gaussians with diagonal covariance to model the observed speech features as suggested by Cooke et al. (2001). Feature vectors for the missing-data recognizer are derived from the DFT coefficients in each frame extracted as described in Section 2. Log compression is applied to the resulting energy spectrum of the signal. To be consistent with experiments on the conventional ASR, we apply a rectangular window to truncate the log-spectral energy to 4 kHz. Hence 98 spectral coefficients along with delta coefficients in a two-frame delta window are extracted in each frame. Note that the delta window covers two preceding frames and two succeeding frames (Young et al., 2000). Additionally, during decoding of the noisy input, the missing-data recognizer uses the estimated binary mask to provide the reliability information for the static features. For the delta features, the binary mask is defined as follows. A unit in the mask is labeled 1 if all the static spectral coefficients used in the computation of the corresponding delta feature are reliable, and 0 otherwise (Barker et al., 2000).

5. Evaluation results

To compare the effect of vocabulary size on the two recognition approaches outlined above, we choose two task domains. The first task is speaker-independent recognition of connected digits. The grammar for this task allows for the repetition of one or more digits. This is the same task used in the original study of Cooke et al. (2001). Thirteen (1–9, a silence, very short pause between words, zero and oh) word-level models are trained for both recognizers. All except the short pause model have 8 emitting states. The short pause model has a single emitting state, tied to state 4 of the silence model. The output distribution in each state is modeled as a mixture of 10 Gaussians, as suggested by Cooke et al. (2001). The grammar for this task allows for one or more repetitions of digits. All digits are equally probable. The average number of digits in an utterance is 3.2, with a minimum of 1 and a maximum of 7. The TIDigits database's male speaker data are used for both training and testing (Leonard, 1984). Specifically, the models are trained using 4235 utterances in the training set of this database. Testing is performed on a subset of the testing set consisting of 461 utterances from 6 speakers, comprising 1498 words. All test speakers are different from the speakers in the training set. The signals in this database are sampled at 20 kHz.

The second task is the speaker-independent recognition of command and control type phrases. There are 40 phrase templates that range from a 1 word template such as “STOP” to a 8 word template such as “CALL MY DAUGHTER AT ELEVEN PM ON <WEEKDAY>”. Note that some templates generate many utterances by using variables such as <WEEKDAY>. The grammar for this task assigns equal probability to all 40 phrase templates in the database. The average number of words in an utterance is 4.2. Two hundred and eight (206 words, a silence and a short pause between words) word-level models are trained for both recognizers. This task allows us to increase the vocabulary size from 13 to 208, a natural progression in testing the effect of vocabulary size on the recognizers. All except the short pause model have 8 emitting states, whose output distribution is modeled as a mixture of 10 Gaussians. The short pause model has a single state. The digital data subset of the Apple Words and Phrases database is used for both training and testing (Cole et al., 1995). In particular, 1996 speakers with IDs 21–2604 are used for training. This corresponds to 63,835 utterances. Data from 14 speakers with IDs 4–19 are used for testing. This corresponds to 454 utterances, comprising 1823 words. The signals are sampled at 8 kHz.

The two tasks also differ in perplexity. Perplexity is one indicator of difficulty of the recognition task along with vocabulary size (Rabiner and Juang, 1993). Perplexity is a measure of the average number of different words that can occur following any given word. For the digit recognition task, the perplexity is 11.0 as any digit can follow any other digit. For the command and control task the perplexity is 3.05. For our task, we calculate the perplexity empirically from the word level lattice (Young et al., 2000). The lower perplexity for the second task is due to the use of a restrictive grammar for this task (Cole et al., 1995). To test the robustness of the two recognizers in the aforementioned tasks, noise is added at a range of SNRs from -5 to 10 dB in steps of 5 dB. Higher positive values of SNRs are not explored, as one of the recognizers saturates to ceiling performance at 10 dB. The noise source for both recognition tasks is the factory noise from the NOISEX corpus (Varga et al., 1992), which is also used by Cooke et al. (2001). The factory noise is chosen as it has energy in the formant regions, therefore posing challenging problems for recognition. It also contains contributions from impulsive sources, making it difficult to estimate its spectrum using spectral subtraction methods (Cooke et al., 2001). In all our experiments, the target speech source is in the median plane and the noise source on the right side at 30°, making the left ear the better ear in terms of SNR (see Section 3). The binaural mixture is created by convolving the monaural target and noise signals with the HRTFs corresponding to the respective source locations as described in Section 2. Note that for both tasks, training and testing are performed using speech band-limited to 4 kHz.

Fig. 4 summarizes the performance of the two recognizers on the digit recognition task. Performance is measured in terms of word-level recognition accuracy under various SNR conditions. “Unprocessed” refers to the baseline performance of the conventional ASR, without the use of any front-end, speech enhancement processing. The figure shows the recognition accuracy of the conventional ASR with the use of ideal and estimated ratio T-F masks (“Ideal RM” and “Estimated RM”, respectively). This is compared to the accuracy of the missing-data recognizer, which uses ideal and estimated binary T-F masks (“Ideal BM” and “Estimated BM”, respectively). We also compare the performance to that obtained by using an advanced front-end feature extraction algorithm (“ETSI AFE”), which is standardized by the European Telecommunication Standards Institute (ETSI) (STQ-AURORA, 2005). This is a state-of-the-art feature extraction algorithm for noisy environments (Macho et al., 2002). The algorithm achieves noise robustness by combining a two-stage Wiener filter with an SNR-dependent waveform processing. Finally, power spectra extraction and blind equalization are used in the cepstrum calculation in order to further improve robustness of the extracted features. The algorithm is used to generate a 39 dimensional feature vector as suggested by Macho et al. (2002). These features are used to train and test an ASR system in a manner similar to that used by the conventional ASR.

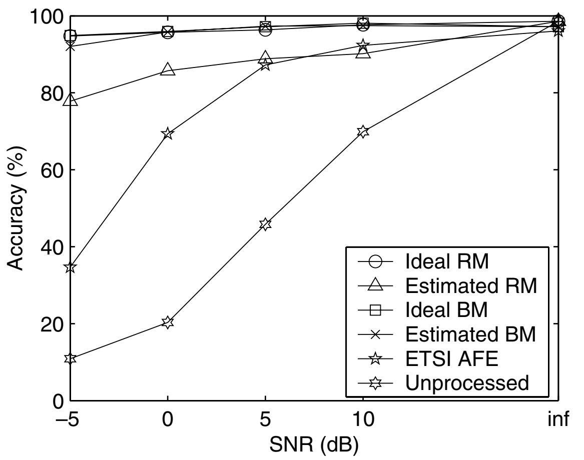

Fig. 4. Performance of conventional and missing-data recognizers on the digits recognition task. Ideal RM refers to the performance of the conventional ASR using the ideal ratio mask. Estimated RM refers to its performance when using the estimated ratio mask by the binaural front-end. Ideal BM refers to the performance of the missing-data ASR using the ideal binary mask. Estimated BM refers to the performance of the same when using the estimated binary mask by the binaural front-end. For comparison, the performance of the conventional ASR without the use of any front-end processing and with processing by the ETSI advanced feature extraction algorithm are also shown.

Fig. 4 shows the robust performance of the ideal ratio mask when used as a front-end for conventional ASR. Only a minor performance degradation is observed even at -5 dB. The performance with the use of the estimated ratio mask shows substantial improvement over that with no preprocessing across all SNR conditions. Additionally, the estimated ratio mask also outperforms the ETSI advanced front-end at SNR <10 dB. The performance improvement is especially substantial at low SNRs. As reported by Cooke et al. (2001), the performance of the missing-data recognizer degrades very little with increasing amounts of noise added, indicating the adequacy of recognition using a binary mask for this task. Also, the performance with the estimated binary mask is close to that with the ideal binary mask, indicating the high quality of the front-end to estimate the ideal binary mask (see also Roman et al., 2003). Notice that, for this task, the performance of the missing-data recognizer is close to the performance of the conventional ASR with the ideal ratio mask and better than the performance with the estimated ratio mask.

Similarly, Fig. 5 summarizes the performance of the two recognizers on the task of recognition of command and control phrases. The relative performance of the two recognizers reverses with this increase in the vocabulary size. As in the digits recognition task, the performance of the conventional ASR using the ideal ratio mask is close to the ceiling performance. Additionally, its performance using the estimated ratio mask is close to the that with the ideal ratio mask, especially at SNR >0 dB. Also notice that the estimated ratio mask outperforms the ETSI advanced feature extraction algorithm across all SNRs. As in the previous task, the performance improvement is substantial at low SNRs. The increased accuracy of the conventional ASR using the estimated ratio mask compared to its performance on the digits recognition task is due to the lower perplexity of this task. Its performance now is substantially better than that of the missing-data recognizer using both ideal and estimated binary masks, particularly at SNR $\geqslant 0$ dB. Notice that the performance of the missing-data recognizer with the estimated binary mask is close to that with the ideal binary mask as in the digits recognition task, confirming the ability of the front-end to estimate the ideal binary mask accurately.

In the experiments with the missing-data ASR thus far, the threshold θ used in defining the ideal binary mask (Eq. (2)) is fixed at 0.5. Fig. 6 shows the performance of the missing-data ASR on both tasks as θ is varied in the range 0.2–0.8. The performance is measured in terms of change in recognition accuracy relative to that obtained by setting $\theta = 0.5$. The SNR is 0 dB. Note that for both tasks, the performance is relatively stable for θ in the range 0.4–0.7. While the highest accuracy on the digit recognition task is obtained when $\theta = 0.3$, the improvement is only small. Moreover, this choice of threshold results in a significant degradation in the performance on the command and control task. For this task, $\theta = 0.5$ provides the best performance. Hence, the selection of 0.5 as the threshold in the definition of the ideal binary mask is an appropriate choice for the tasks considered in the present study.

Lower accuracy values for the missing-data recognizer using both binary masks in Fig. 5 may

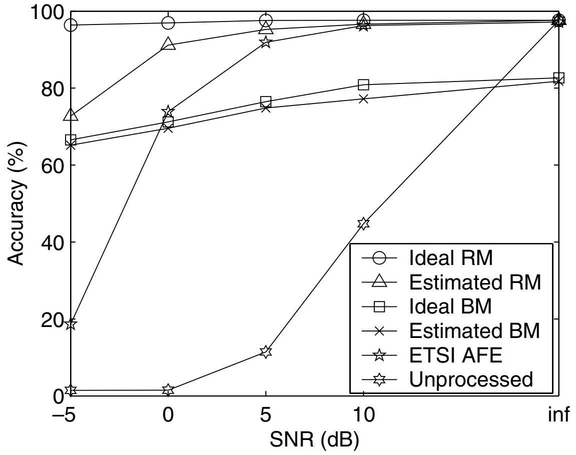

Fig. 5. Performance of conventional and missing-data recognizers on the command and control task. See Fig. 4 caption for notations.

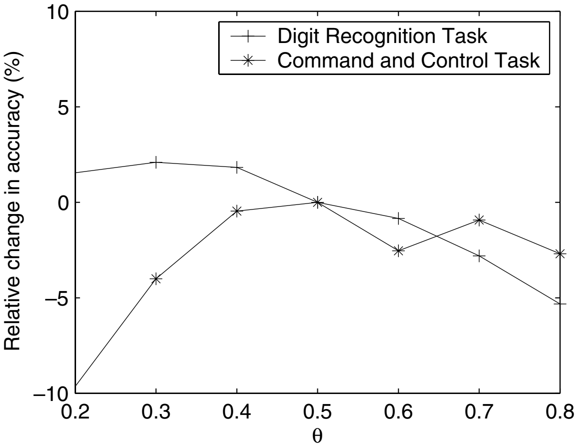

Fig. 6. Effect of varying the threshold, θ, used in the definition of the ideal binary mask (see Eq. (2)) on the performance of the missing-data ASR on the two tasks. The performance is relative to its performance when using $\theta = 0.5$. The SNR for both tasks is 0 dB.

be attributed to a number of reasons. It is known that the use of mixtures of Gaussians with diagonal covariance structure does not adequately represent the observed spectral vectors (Cooke et al., 2001) and this problem gets exacerbated with an increase in the vocabulary size. Thus, under clean speech conditions, the difference between the accuracy of conventional and missing-data recognizers increases with increase in vocabulary size (see also Raj et al., 2004). One could compute MFCCs from the speech resynthesized using the binary T-F masks by substituting the ratio mask with the binary mask in (5), followed by the overlap and add resynthesis (Oppenheim et al., 1999). Under clean speech conditions, the missing-data recognizer would then have the same recognition accuracy as that of the conventional ASR. The performance though degrades rapidly with decreasing SNR (de Veth et al., 1999).

The use of binary masks does not compensate for amplitude distortions, because the mixture spectral values are used in recognition for those T-F units labeled 1. Could this be the reason for reduced performance in larger vocabulary recognition? To test the effect of this distortion, we replace the spectral vectors of the reliable T-F regions with their corresponding clean speech values, calculated *a priori*. The performance, at various SNRs, is summarized in Tables 1 and 2. "Distorted" refers to the performance of the missing-data recognizer on the mixture spectral values for all T-F units. "Undistorted" refers to its performance when the reliable T-F units contain clean speech values. In the unreliable units, we retain the spectral values of noisy speech. We use the ideal binary mask generated at each SNR to provide the reliability information for both conditions. Table 1 shows the effect of amplitude distortion on the digits recognition task. For this task, the effect of amplitude distortion is seen, as expected, to be minimal across all SNRs, since the recognition accuracy is already quite high. Table 2 shows the effect of amplitude distortion on the task of command and control phrases. Only a small improvement is observed by eliminating the noise energy from the reliable T-F units. Hence, the degradation to the overall performance of the missing-data recognizer caused by this amplitude distortion is statistically insignificant at the range of SNRs considered here. When using the ideal binary mask generated at each SNR directly on clean speech, we observe a degradation in performance. This may be attributed to the use of energy bounds for the unreliable units in the marginalization method. The bounded marginalization method averages the observation probability over all possible spectral energy values between 0 and the observed value. Hence, when the observed value is the clean spectral energy, the bounded marginalization method overestimates the true observation probability.

Table 1
Effect of amplitude distortion in the reliable T-F regions on recognition accuracy (%) of the missing-data recognizer for the digits recognition task

Amplitude	SNR (dB)			
	−5	0	5	10
Distorted	94.89	95.97	97.18	98.12
Undistorted	94.76	96.24	97.31	98.25

Table 2
Effect of amplitude distortion in the reliable T-F regions on recognition accuracy (%) of the missing-data recognizer for the command and control task

Amplitude	SNR (dB)			
	−5	0	5	10
Distorted	66.54	71.2	76.51	80.88
Undistorted	69.17	74.38	79.1	82.67

Comparing Figs. 4 and 5, we can see that the performance curve for the missing-data recognizer is steeper on the second task compared to the first task. This may be caused by the inability of the missing-data recognizer to represent all the speech models adequately. The log-spectral representation may have a limited expressibility in terms of distinct words that can be uniquely represented. The TIDigits database has a small vocabulary. The Applewords database with a larger vocabulary creates many more competing models during decoding. Thus, within the same T-F grid, an increased number of words need to be discriminated. With the use of a binary mask, only a small portion of the total T-F acoustic model space is utilized during recognition. This makes it difficult for the missing-data recognizer to differentiate between competing hypotheses. Fig. 7 shows the effect of using the same binary T-F mask on two signals. Fig. 7(a) shows the spectrogram of the word "Billy" and Fig. 7(b) shows the spectrogram of the word "Delete" in quiet. Fig. 7(c) shows an ideal binary T-F mask generated at low SNR. The reliable units in this mask are black and the unreliable white. This binary mask is applied to the spectrograms in Figs. 7(a) and (b) and the resulting spectrograms with only reliable

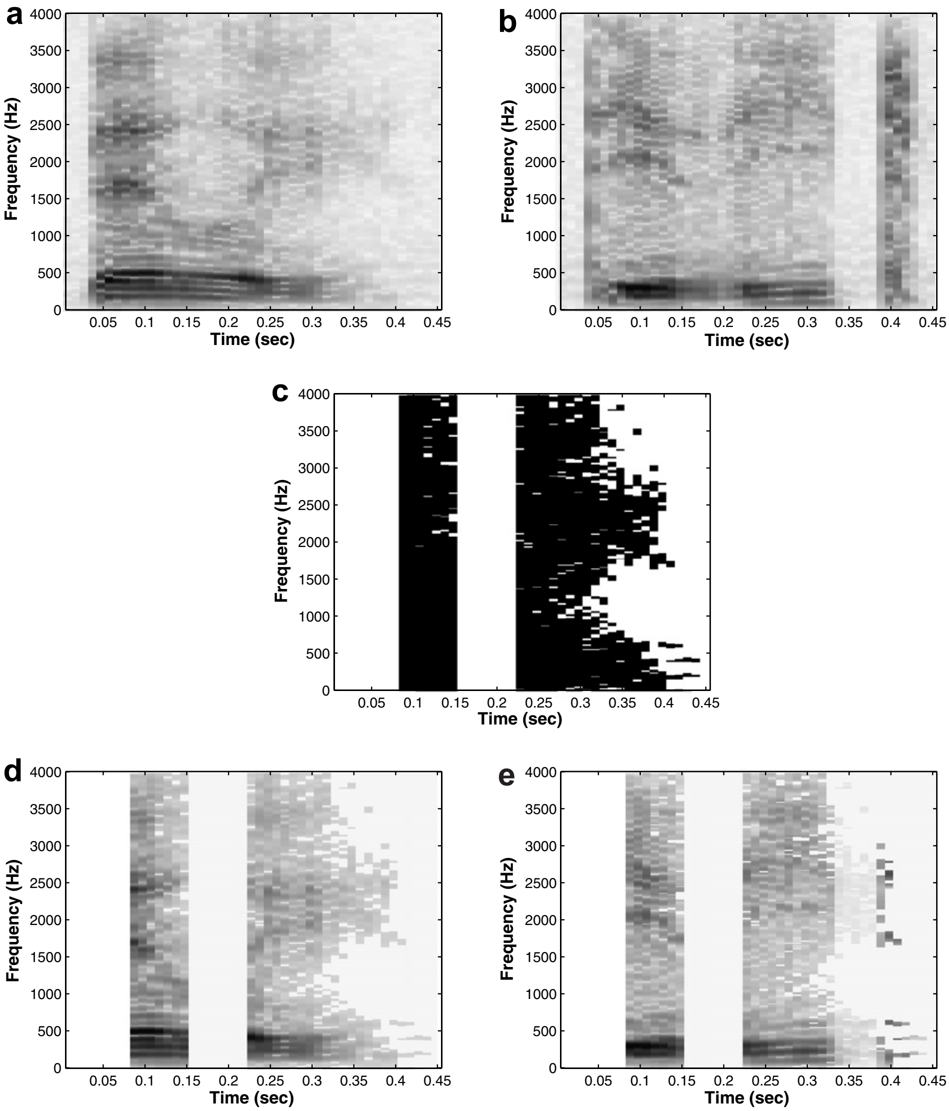

Fig. 7. An illustration of similarity of reliable regions. (a) The spectrogram of the word "Billy" in quiet. (b) The spectrogram of the word "Delete" in quiet. (c) An ideal-binary T-F mask. Reliable T-F units are marked black and unreliable white. (d) The spectrogram obtained from (a) by applying the ideal mask in (c). (e) The spectrogram obtained from (b) by the same ideal masking as in (d).

T-F units are shown in Figs. 7(d) and (e), respectively. Notice that the reliable regions of the two spectrograms are very similar. In the absence of information in the unreliable regions, it is difficult for the recognizer to distinguish between the two words. Indeed the recognizer frequently substitutes one word with the other. The bounded marginalization method treats the information in the unreliable regions only as counter-evidence for recognition of certain models (Cunningham and Cooke, 1999). Hence, the missing-data recognizer faces increased acoustic complexity during decoding.

6. Discussion

The advantage of the missing-data recognizer is that it imposes a lesser demand on the speech enhancement front-end than the conventional ASR. Only knowledge of reliable T-F units of noisy speech, or an ideal binary mask, is required from the front-end. Moreover, Roman et al. (2003) have shown that the performance of the missing-data recognizer degrades gradually with increasing deviation from the ideal binary mask. The binaural system employed here is able to estimate this mask accurately. Hence, we achieve performance close to the ceiling performance of missing-data recognition. Conventional ASR, on the other hand, requires speech enhancement across all T-F units through front-end processing. In this study, we have employed a ratio T-F mask as a front-end for the conventional ASR, which is estimated using statistics of ITD and IID. Estimation of the ideal ratio mask is less robust than the estimation of the ideal binary mask. As a result, the performance of the missing-data recognizer on the small vocabulary task is better than that of the conventional ASR.

The marginalization method for missing-data recognition is the optimal spectral domain recognition strategy provided that the missing T-F units can be ignored for classification (Little and Rubin, 1987). The missing-data recognizer assumes that the unreliable units carry redundant information for speech recognition. This, however, is not always true. For a small vocabulary task, the unreliable units may be safely marginalized for good recognition results. When vocabulary size increases, the acoustic model space becomes densely populated. Under such conditions, good recognition results may not be obtained by completely ignoring the missing T-F units. This may be caused by the inability to represent all the acoustic models adequately using only a small number of reliable T-F units. On the other hand, the ratio T-F mask attempts to recover the speech in the unreliable T-F units for use in recognition (see Eq. (5)). Additionally, under clean speech conditions, recognition accuracy using spectral features is inferior to using cepstral features. The cepstral transformation retains the envelope of speech while removing its excitation source (Rabiner and Juang, 1993). The speech envelope contains most relevant information for recognition. In addition, cepstral features are used for their quasi-orthogonal properties (Shire, 2000). Hence, advantage of the conventional ASR shows when vocabulary size increases.

Raj et al. (2004) have previously reported that conventional ASR with reconstructed missing T-F regions outperforms the missing-data recognizer when tested on the Resource Management database (Price et al., 1988). The missing or unreliable T-F units were reconstructed either using speech clusters or based on their correlations with reliable regions. The speech clusters and the knowledge of correlations between reliable and unreliable T-F units are obtained from the training portion of the Resource Management database. Unlike their system, our estimation of the ideal ratio mask is independent of the signals used in the training and testing of the speech recognizers. Hence, it is applicable even when samples of clean speech are unavailable. Additionally, the accuracy and computational complexity of our ratio mask based system are not dependent on the nature and size of the vocabulary.

Note that our supervised training captures only the location information, and hence our system is not sensitive to the content of sound sources. While we have shown results only for one configuration of target and interference source positions, similar results are expected for other spatial configurations, including those with interferences at multiple locations (see Roman et al., 2003). The study of Roman et al. (2003) also addresses the localization of various sounds sources in a mixture. However, the estimation of the ratio and the binary masks requires training for different configurations of sources and reestimating the mean curves as described in Section 3. Although extrapolation from trained configurations to untrained ones may be possible, this is a limitation that needs to be addressed in future research.

Although our estimated T-F ratio mask provides promising results, other approaches for the estimation of this mask could also be explored; should a parametric curve be suspected, the parameters could be optimized to minimize recognition errors. Future work will also extend to large vocabulary tasks and explore the robustness of the binaural front-end to changes in location and number of noise sources.

To summarize, we have proposed a ratio T-F mask, estimated using a binaural processor, as a front-end for conventional ASR. At two different vocabulary sizes, the use of this mask results in sizable improvement in recognition accuracy at various SNRs when compared to the baseline performance. Additionally, it significantly outperforms the ETSI advanced robust feature extraction algorithm at most SNRs. On the larger vocabulary task, the performance of the proposed ASR is substantially better than that of the missing-data recognizer. Our study indicates that optimal preprocessing strategies for robust speech recognition may depend on the vocabulary size of the task. For small vocabulary applications, computation of the ideal T-F binary mask may be desirable, whereas a ratio mask may provide an improved performance with increased vocabulary sizes.

Acknowledgments

This research was supported in part by an AFOSR Grant (FA9550-04-1-0117) and an NSF Grant (IIS-0081058). We thank M. Cooke for discussion and assistance in implementing the missing-data recognizer, and D. Pearce for helping us obtain the ETSI advanced front-end algorithm. We also thank the three anonymous reviewers for their suggestions/criticisms. A preliminary version of this work was presented in 2004 ICSLP.

References

Barker, J., Josifovski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. International Conference on Spoken Language Processing '00, pp. 373–376.

Blauert, J., 1997. Spatial Hearing – The Psychophysics of Human Sound Localization. MIT Press, Cambridge, MA.

Boll, S.F., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Processing ASSP-27 (2), 113–120.

Bradstein, M., Ward, D. (Eds.), 2001. Microphone Arrays: Signal Processing Techniques and Applications. Springer, Berlin, Germany.

Brown, G.J., Wang, D.L., 2005. Separation of speech by computational auditory scene analysis. In: Benesty, J., Makino, S., Chen, J. (Eds.), Speech Enhancement. Springer, New York, pp. 371–402.

Cardoso, J.F., 1998. Blind signal separation: statistical principles. Proc. IEEE 86 (10), 2009–2025.

Chen, C.-P., Bilmes, J., Ellis, D.P.W., 2005. Speech feature smoothing for robust ASR. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing '05, vol. 1, pp. 525–528.

Cole, R., Noel, M., Lander, T., Durham, T., 1995. New telephone speech corpora at CSLU. In: Proc. European Conference on Speech Communication and Technology '95, pp. 821–824.

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

Cunningham, S., Cooke, M., 1999. The role of evidence and counter-evidence in speech perception. In: Proc. International Congress on Phonetic Sciences '99, pp. 215–218.

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

de Veth, J., de Wet, F., Cranen, B., Boves, L., 1999. Missing feature theory in ASR: make sure you miss the right type of features. In: Proc. Workshop on Robust Methods for Speech Recognition in Adverse Conditions '99, pp. 231–234.

Droppo, J., Acero, A., Deng, L., 2002. A nonlinear observation model for removing noise from corrupted speech log mel-spectral energies. In: Proc. International Conference on Spoken Language Processing '02, pp. 1569–1572.

Ehlers, F., Schuster, H.G., 1997. Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment. IEEE Trans. Signal Processing 45 (10), 2608–2612.

Ephraim, Y., 1992. A Bayesian estimation approach for speech enhancement using hidden Markov models. IEEE Trans. Signal Processing 40 (4), 725–735.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Processing ASSP-32 (6), 1109–1121.

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

Gardner, W.G., Martin, K.D., 1994. HRTF measurements of a KEMAR dummy-head microphone. Technical Report #280, MIT Media Lab Perceptual Computing Group.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, N., 1993. Darpa timit acoustic–phonetic continuous speech corpus. Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD.

Glotin, H., Berthommier, F., Tessier, E., 1999. A CASA-labelling model using the localisation cue for robust cocktail-party speech recognition. In: Proc. European Conference on Speech Communication and Technology '99, pp. 2351–2354.

Gong, Y., 1995. Speech recognition in noisy environments: a survey. Speech Commun. 16, 261–291.

Hughes, T.B., Kim, H.S., DiBase, J.H., Silverman, H.F., 1999. Performance of an HMM speech recognizer using a real-time tracking microphone array as input. IEEE Trans. Speech Audio Processing 7 (3), 346–349.

Leonard, R.G., 1984. A database for speaker-independent digit recognition. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing '84, pp. 111–114.

Lippmann, R.P., 1997. Speech recognition by machines and humans. Speech Commun. 22, 1–15.

Little, R.J.A., Rubin, D.B., 1987. Statistical Analysis with Missing Data. Wiley, New York, NY.

Macho, D., Mauuary, L., Noe, B., Cheng, Y.M., Ealey, D., Jouvet, D., Kelleher, H., Pearce, D., Saadoun, F., 2002. Evaluation of a noise-robust DSR front-end on aurora databases. In: Proc. International Conference on Spoken Language Processing '02, pp. 17–20.

McAulay, R., Malpass, M.L., 1980. Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Processing ASSP-28 (2), 137–145.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-time Signal Processing. second ed. Prentice-Hall, Upper Saddle River, NJ.

Palomaki, K.J., Brown, G.J., Wang, D.L., 2004. A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation. Speech Commun. 43, 361–378.

Price, P., Fisher, W.M., Bernstein, J., Pallett, D.S., 1988. The DARPA 1000 word Resource Management database for continuous speech recognition. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing '88, pp. 651–654.

Rabiner, L.R., Juang, B.H., 1993. Fundamentals of Speech Recognition. second ed. Prentice-Hall, Englewood Cliffs, NJ.

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

Roman, N., Wang, D.L., Brown, G.J., 2003. Speech segregation based on sound localization. J. Acoust. Soc. Am. 114, 2236–2252.

Rosenthal, D.F., Okuno, H.G. (Eds.), 1998. Computational Auditory Scene Analysis. Lawrence Erlbaum Associates, Mahwah, NJ.

Shire, M.L., 2000. Discriminant training of front-end and acoustic modeling stages to heterogeneous acoustic environments for multi-stream automatic speech recognition. Ph.D. thesis, University of California, Berkley.

Srinivasan, S., Roman, N., Wang, D.L., 2004. On binary and ratio time-frequency masks for robust speech recognition. In: Proc. International Conference on Spoken Language Processing '04, pp. 2541–2544.

STQ-AURORA, 2005-11. Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms. In: ETSI ES 202 050 V1.1.4. European Telecommunications Standards Institute.

Tessier, E., Berthommier, F., Glotin, H., Choi, S., 1999. A CASA front-end using the localisation cue for segregation and then cocktail-party speech recognition. In: Proc. IEEE International Conference on Speech Processing, pp. 97–102.

van Hamme, H., 2003. Robust speech recognition using missing feature theory in the cepstral or LDA domain. In: Proceedings of the European Conference on Speech Communication and Technology '03, pp. 3089–3092.

van Trees, H.L., 1968. Detection, Estimation, and Modulation Theory, Part I. Wiley, New York, NY.

Varga, A.P., Moore, R.K., 1990. Hidden Markov model decomposition of speech and noise. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing '90, pp. 845–848.

Varga, A.P., Steeneken, H.J.M., Tomlinson, M., Jones, D., 1992. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK.

Young, S., Kershaw, D., Odell, J., Valtchev, V., Woodland, P., 2000. The HTK Book (for HTK Version 3.0). Microsoft Corporation.