Technical Report OSU-CISRC-9/05-TR60
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftp site: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2005**
File: **TR60.pdf**
Web site: **http://www.cse.ohio-state.edu/research/tech-report.html**

Binaural Segregation in Multisource Reverberant Environments

Nicoleta Roman
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
niki@cse.ohio-state.edu

Soundararajan Srinivasan
Biomedical Engineering Center
The Ohio State University, Columbus, OH 43210, USA
srinivasan.36@osu.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive Science
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

Correspondence should be directed to D. Wang: Department of Computer Science and Engineering, The Ohio State University, 2015 Neil Avenue, Columbus, OH 43210. Phone: (614)-292-6827, URL: www.cis.ohio-state.edu/~dwang.

ABSTRACT

In a natural environment, speech signals are degraded by both reverberation and concurrent noise sources. While human listening is robust under these conditions using only two ears, current two-microphone algorithms perform poorly. The psychological process of figure-ground segregation suggests that the target signal is perceived as foreground while the remaining stimuli are perceived as background. Accordingly, our goal is to estimate an ideal time-frequency (T-F) binary mask, which selects the target if it is stronger than the interference in a local T-F unit. In this paper, we propose a binaural segregation system which extracts the reverberant target signal from multisource reverberant mixtures by utilizing only the location information of target source. The proposed system combines target cancellation through adaptive filtering and a binary decision rule to estimate the ideal T-F binary mask. A key observation in this work is that the attenuation due to target cancellation in a T-F unit is systematically correlated with the relative strength between target and interference. A comprehensive evaluation shows that the proposed system results in large SNR gains. In addition, comparisons using SNR as well as automatic speech recognition measures show that our system outperforms standard two-microphone beamforming approaches and a recent binaural processor.

I. INTRODUCTION

A typical auditory environment contains multiple concurrent sources that are reflected by surfaces and change their locations constantly. While human listeners are able to attend to a particular sound signal even under such adverse conditions, simulating this perceptual ability or solving the cocktail party problem (Cherry, 1953) remains a grand challenge. A solution to the problem of sound separation in real environments is essential for many applications including automatic speech recognition (ASR), audio information retrieval and hearing prosthesis. In this paper we study the binaural (two-microphone) separation of speech in multisource reverberant environments.

The sound separation problem has been investigated in the signal processing field for many years for both one-microphone recordings as well as multi-microphone ones (for recent reviews see Divenyi, 2005; Brandstein and Ward, 2001). One-microphone speech enhancement techniques include spectral subtraction (e.g., Martin, 2001), Kalman filtering (Ma et al., 2004), subspace analysis (Ephraim and Trees, 1995) and autoregressive modeling (e.g., Balan et al., 1999). While having the advantage of requiring only one sensor, these algorithms make strong assumptions about environment and thus have difficulty in dealing with general acoustic mixtures. Microphone array algorithms are divided in two broad categories: beamforming and independent component analysis (ICA) (Brandstein and Ward, 2001). To separate multiple sound sources, beamforming takes advantage of their different directions of arrival while ICA relies on their statistical independence. A fixed beamformer, such as that of the delay-and-sum, constructs a spatial beam to enhance signals arriving from the target direction independent of the interfering sources. The primary limitations of a fixed beamfomer are: 1) a poor spatial resolution at lower frequencies, i.e., the spatial response has a wide main lobe when the intermicrophone distance is smaller than the signal wavelength; and 2) spatial aliasing, i.e., multiple beams at higher frequencies when the intermicrophone distance is greater than the signal wavelength. To solve these problems a large number of microphones is required and constraints need to be introduced in order to impose a constant beam shape across the frequencies (Ward et al., 2001). Adaptive beamforming techniques, on the other hand, attempt to null out the interfering sources in the mixture (Griffiths and Jim, 1982; Widrow and Stearns, 1985; Van Compernolle, 1990). While they improve spatial resolution significantly, the main disadvantage of such beamformers is greater computation and adaptation time when the locations of interfering sources change. Note also that while an adaptive beamformer with two microphones is optimal for canceling a single directional interference, additional microphones are required as the number of noise sources increases (Weiss, 1987). A subband adaptive algorithm has been proposed by Liu et al. (2001) to address the multi-source problem. Their two-microphone system estimates the locations of all the interfering sources and uses them to steer independent nulls that suppress the strongest interference in each T-F unit. The underlying signal model is, however, anechoic and performance degrades in reverberant conditions. Similarly, the drawbacks of ICA techniques include the requirement that the number of microphones be greater than or equal to the number of sources and poor performance in reverberant conditions (Hyväarinen et al., 2001). Some recent sparse representations attempt to relax the former assumption (e.g., Zibulevsky et al., 2001) but the performance is limited.

However, human listeners excel at separating target speech from multiple interferences. Inspired by this robust performance, research has been devoted to build speech separation systems that incorporate known principles of auditory perception. According to Bregman (1990), the auditory system performs sound separation by employing various grouping cues including pitch, onset time, spectral continuity and location in a process known as auditory scene analysis (ASA). This ASA account has inspired a series of computational ASA (CASA) systems that have significantly advanced the state-of-the-art performance in monaural separation as well as in binaural separation. Monaural separation algorithms rely primarily on the pitch cue and therefore operate only on voiced speech. On the other hand, the binaural algorithms use the source location cues - time differences and intensity differences between the ears – which are independent of the signal content and thus can be used to track both voiced and unvoiced speech. A recent overview of CASA approaches can be found in Brown and Wang (2005).

CASA research, however, has been largely limited to anechoic conditions, and few systems have been designed to operate on reverberant inputs. In reverberant conditions, anechoic modeling of time delayed and attenuated mixtures is inadequate. Reverberation introduces potentially an infinite number of sources due to reflections from hard surfaces. As a result, the estimation of location cues in individual T-F units becomes unreliable with an increase in reverberation and the performance of location-based segregation systems degrade under these conditions. A notable exception is the binaural system proposed by Palomäki et al. (2004) which includes an inhibition mechanism that emphasizes the onset portions of the signal and groups them according to common location. The system shows improved speech recognition results across a range of reverberation times. Evaluations in reverberation have also been reported for two-microphone algorithms that combine pitch information with binaural cues or other signal-processing techniques (Luo and Denbigh, 1994; Shamsoddini and Denbigh, 1999; Barros et al., 2002).

From an information processing perspective, the notion of an *ideal T-F binary mask* has been proposed as the computational goal of CASA (Roman et al., 2003; see also Wang, 2005). Such a mask can be constructed from *a priori* knowledge of target and interference; specifically a value of 1 in the mask indicates that the target is stronger than the interference within a particular T-F unit and 0 indicates otherwise. Speech reconstructed from ideal binary masks has been shown to be highly intelligible even when extracted from multi-source mixtures and also produce substantial improvements in robust speech recognition (Cooke et al., 2001; Roman et al., 2003; Brungart et al., 2005).

As stated earlier, only one source can be canceled through linear filtering in binaural processing. In this paper we pursue a binaural solution to target segregation under reverberant conditions and in the presence of multiple concurrent sound sources. We propose a two-stage model that combines target cancellation followed by a nonlinear processing stage to estimate the ideal binary mask. Specifically, we achieve target cancellation through adaptive filtering and observe a correlation between the amount of cancellation produced in individual T-F units and the relative strength between target and interference. Consequently, the input-to-output attenuation level is employed to estimate the ideal binary mask. Since the system depends only on the location of the target, it works for a variety of interfering sources including moving intrusions and impulsive ones. Álvarez et al. (2002) proposed a related system that combines a first-order differential beamformer to cancel the target and obtain a noise estimate and spectral subtraction to enhance the target source, but their results are not satisfactory in reverberant conditions.

Although the speech reconstructed directly from the ideal binary mask is highly intelligible, typical ASR systems are sensitive to the small distortions produced during resynthesis and hence do not perform well on the reconstructed signals. Two methods have been proposed to alleviate this problem: 1) the missing-data ASR proposed by Cooke et al. (2001) which utilizes only the reliable (target dominant) features in the acoustic mixture; and 2) a target reconstruction method for the unreliable (interference dominant) features proposed by Raj et al. (2004) followed by a standard ASR system. While the first method constraints the ASR to operate on spectral features, the second method reconstructs the spectrograms during front-end processing allowing the ASR to capitalize on the advantage of cepstral features. In our evaluations, we use a spectrogram reconstruction technique similar to the one proposed by Raj et al. (2004) and show substantial speech recognition improvements over baseline and other related two-microphone approaches.

The rest of the paper is organized as follows. The next section defines the problem and describes the model. Section 3 gives an extensive evaluation of our system as well as a comparison with related models. The last section concludes the paper.

II. MODEL ARCHITECTURE

The proposed model consists of two stages as shown in Fig. 1. In the first stage, the system performs target cancellation through adaptive filtering. In the second stage, the system labels as 1 those T-F units that have been largely attenuated in the first stage since those units are likely to have originated from the target source.

The input signal shown in Fig. 1 assumes that a desired speech source s has been produced in a reverberant enclosure and recorded by two microphones to produce the signal pair (x_1, x_2). The transmission path from target location to microphones is a linear system and is modeled as:

$$x_1(t) = h_1(t) * s(t)\,, \tag{1a}$$

$$x_2(t) = h_2(t) * s(t)\,, \tag{1b}$$

where h_i corresponds to the room impulse response for the i'th microphone. The challenge of source separation arises when an unwanted interference pair (n_1, n_2) is also present at the input of the microphones. The interference here is a combination of multiple reverberant sources and additional background noise. In this study, the target is assumed to be fixed but no restrictions are imposed on the number, location, or content of the interfering sources. In realistic conditions, the interference can suddenly change its location and may also contain impulsive sounds. Under these conditions, it is hard to localize each individual source in the scene. The goal is therefore to remove or attenuate the noisy background and recover the reverberant target speech based only on the target source location.

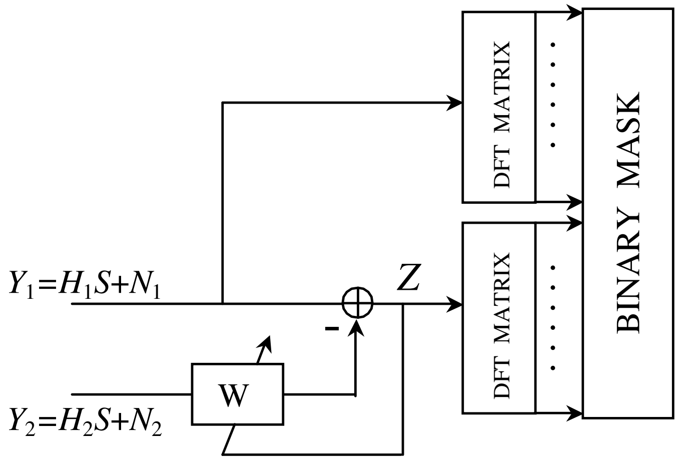

Figure 1. Schematic diagram of the proposed model. The input signal is a mixture of reverberant target sound and acoustic interference. At the core of the system is an adaptive filter for target cancellation. The output of the system is an estimate of the ideal binary mask.

Our objective here is to develop an effective mechanism to estimate an ideal binary mask, which selects the T-F units where the local SNR exceeds a threshold of 0 dB. The relative strength between target signal and interference for a T-F unit is defined as:

$$R(\omega,t) = \frac{|X_1(\omega,t)|}{|X_1(\omega,t)| + |N_1(\omega,t)|}\,, \tag{2}$$

where $X_1(\omega,t)$ and $N_1(\omega,t)$ are the corresponding Fourier transforms of the reverberant target signal and the noise signal at frequency ω and time t corresponding to microphone 1 (primary microphone). Note that the noise signal includes all the interfering sources. Thus, a T-F unit is set to 1 in the ideal binary mask if $R(\omega,t)$ exceeds 0.5, otherwise it is set to 0.

In the classical adaptive beamforming approach (Griffith and Jim, 1982), the filter learns to identify the differential acoustic transfer function of a particular noise source and thus perfectly cancels only one directional noise source. Systems of this type, however, are unable to cope well with multiple noise sources or diffuse background noise. As an alternative, we propose to use the adaptive filter only for target cancellation and then process the noise reference obtained using a nonlinear scheme described below in order to obtain an estimate of the ideal binary mask (see also Roman and Wang, 2004). This approach offers a potential solution to the problem of multiple interfering sources in the background.

In the experiments reported here, we assume a fixed target location and the filter in the target cancellation module (TCM) is trained in the absence of interference. A white noise sequence of 10 s duration is used to calibrate the filter. We implement the adaptation using the Fast-Block Least Mean Square algorithm with an impulse response of 375 ms length (6000 samples at 16 kHz sampling rate) (Haykin, 2002). After the training phase, the filters parameters are fixed and the system is allowed to operate in the presence of interference. Both the TCM output $Z(\omega, t)$ and the noisy mixture at the primary microphone $Y_1(\omega, t)$ are analyzed using a short time-frequency analysis. The time-frequency resolution is 20-ms time frames with a 10-ms frame shift and 257 discrete Fourier transform coefficients. Frames are extracted by applying a running Hamming window to the signal.

As a measure of signal suppression at the output of the TCM unit, we define the output-to-input energy ratio as follows:

$$OIR(\omega,t) = \frac{|Z(\omega,t)|^2}{|Y_1(\omega,t)|^2}, \tag{3}$$

Consider a T-F unit in which noise is zero. Ideally, the TCM module cancels perfectly the target source resulting in zero output and therefore $OIR(\omega,t) \to 0$. On the other hand, T-F units dominated by noise are not suppressed by the TCM and thus $OIR(\omega,t) >> 0$. Hence, a simple binary decision can be implemented by imposing a decision threshold on the estimated output-to-input energy ratio. The estimated binary mask is 1 in those T-F units where $OIR(\omega,t) > \theta(\omega)$ and 0 in all the other units.

Figure 2 shows a scatter plot of *R* and *OIR* obtained for individual T-F units corresponding to a frequency bin at 1 kHz. The results are extracted from 100 mixtures of reverberant target speech fixed at 0° azimuth mixed with four interfering speakers at -135°, -45°, 45° and 135° azimuths. The room reverberation time, T_{60}, is 0.3 s; T_{60} is the time required for the sound level to drop by 60 dB following the sound offset. The input SNR considering reverberant target as signal is 5 dB. Observe that there exists a correlation between the amount of cancellation in the individual T-F units and the relative strength between target and interference. In order to simplify the estimation of the ideal binary mask we have used in our evaluations a frequency-independent threshold of –6 dB on the output-to-input energy ratio. The -6 dB threshold is obtained when the reverberant target signal and the noise have equal energy in Eq. 2. As seen in the figure, the binary masks estimated using this threshold remove most of the noise at the expense of some target speech energy loss. However, informal listening experiments show that the reconstructed signals remain highly intelligible.

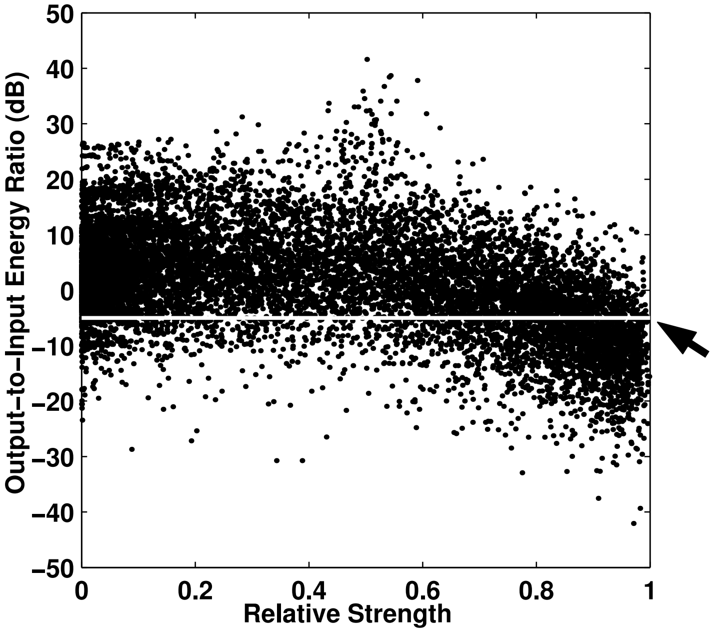

Figure 2. Scatter plot of the output-to-input attenuation with respect to the relative strength for a frequency bin centered at 1 kHz. The white line corresponds to the –6 dB decision threshold used in the binary mask estimation.

III. EVALUATION AND COMPARISON

We have evaluated our system on binaural stimuli, simulated using the room acoustic model described in Palomäki et al. (2004). The reflection paths of a particular sound source are obtained using the image reverberation model for a small rectangular room (6m×4m×3m) (Allen and Berkley, 1979). The resulting impulse response is convolved with the measured head related impulse responses (HRIR) (Gardner et al., 1994) of a KEMAR dummy head (Burkhard and Sachs, 1975) in order to produce the two binaural inputs to our system. Specific room reverberation times are obtained by varying the absorption characteristics of room boundaries. The position of the listener was fixed asymmetrically at (2.5m×2.5m×2m) to avoid obtaining near identical impulse responses at the two microphones when the source is in the median plane. All sound sources are presented at different angles at a distance of 1.5 m from the listener. For all our tests, target is fixed at 0° azimuth unless otherwise specified. To test the robustness of the system to various noise configurations we have performed the following tests : 1) an interference of rock music at 45° (Scene 1); 2) two concurrent speakers (one female and one male utterance) at azimuth angles of -45° and 45° (Scene 2); and 3) four concurrent speakers (two female and two male utterances) at azimuth angles of -135°, -45°, 45° and 135° (Scene 3). The initial and the last speech pauses in the interfering utterances have been deleted in Scene 2 and Scene 3 making them more comparable with Scene 1. The signals are upsampled to the HRIR sampling frequency of 44.1 kHz and convolved with the corresponding left and right ear HRIRs to simulate the individual sources for the above three testing conditions (Scene 1-Scene 3). Finally, the reverberated signals at each ear are summed and then downsampled to 16 kHz. In all our evaluations, the input SNR is calculated at the left ear using reverberant target speech as signal. While in Scene 2 and Scene 3 the SNR at the two ears is comparable, the left ear is the‘better ear’ – the ear with higher SNR - in the Scene 1 condition. In the case of multiple interferences, the interfering signals are scaled to have equal energy at the left ear.

The binaural input is processed by our system as described in Section II in order to estimate the ideal T-F binary mask which is defined as 1 when the reverberant target energy is greater than the interference energy and 0 otherwise. In all our results, the binary mask is computed and the signal is resynthesized at the better ear (left ear). Figure 3 demonstrates the performance of our system for Scene 3 where target utterance is the male utterance “Bright sunshine shimmers on the ocean”. The room conditions are T_{60}=0.3 s and 5 dB input SNR. Observe that the estimated mask is able to estimate well the ideal binary mask especially in the high target energy T-F regions.

To systematically evaluate our segregation system, we use the following performance measures: 1) SNR evaluation using the reverberant target speech as signal; and 2) ASR accuracy using our model as front-end. Quantitative comparisons with related approaches are also provided.

A. SNR evaluation

We perform SNR evaluations for the three conditions described above using 10 speech signals from the TIMIT database (Garofolo et al., 1993) as target: five female utterances and five male utterances. Results are given in Table 1, Table 2 and Table 3. The room reverberation time is 0.3 s in all conditions and the system is evaluated for the following four input SNR values: -5 dB, 0 dB, 5 dB and 10 dB. In order to assess the system performance, output SNR and retained speech ratio (RSR) are computed as follows:

$$Output\ SNR = 10\log 10\left(\sum_t s_E{}^2(t) \Big/ \sum_t n_E{}^2(t) \right) \tag{4}$$

$$RSR = \sum_t s_E{}^2(t) \Big/ \sum_t s_T{}^2(t) \tag{5}$$

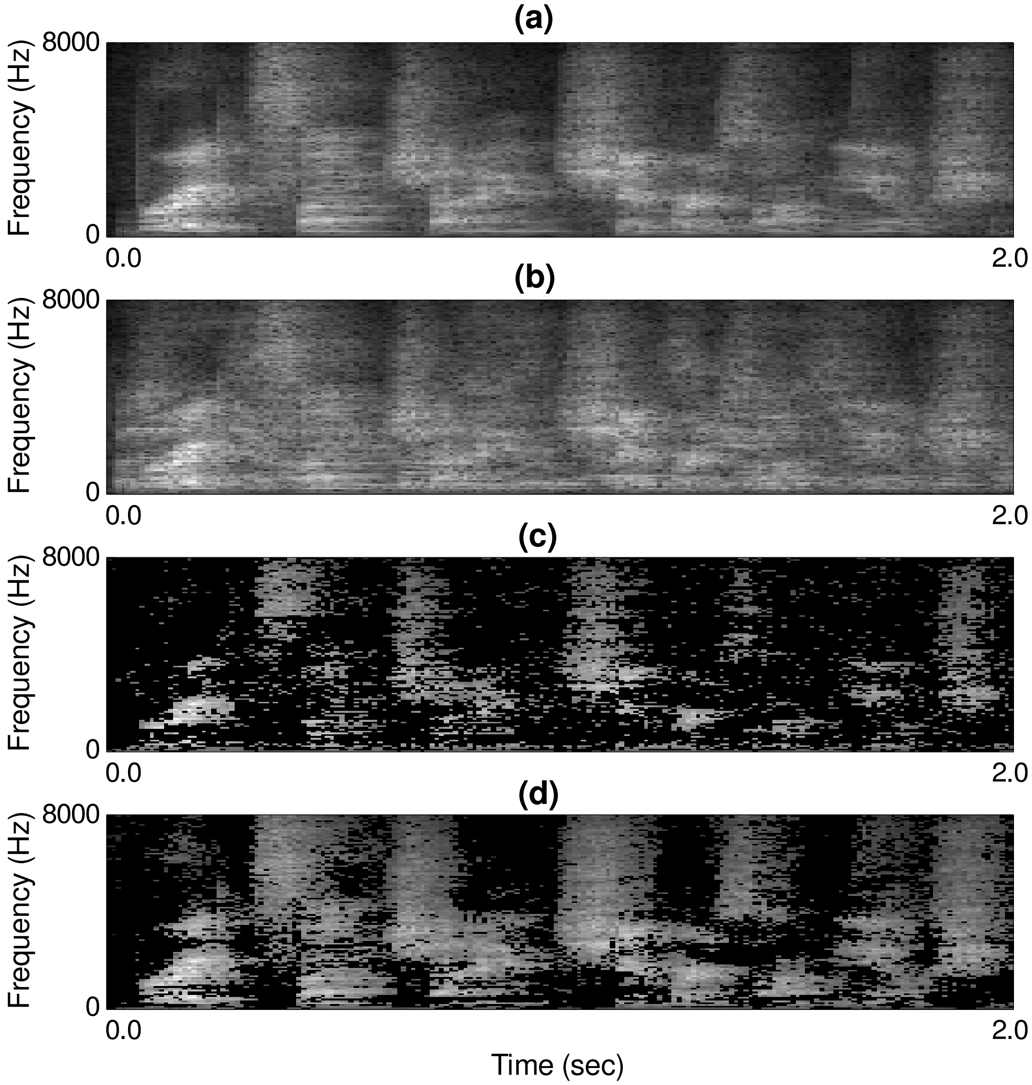

Figure 3. Comparison between the estimated mask and the ideal binary mask for a five-source configuration. (a) Reverberant target speech. (b) Mixture of target speech presented at 0° and four interfering speakers at locations -135°, -45°, 45° and 135°. The SNR is 5 dB. (c) The mixture spectrogram overlaid by the estimated T-F binary mask. (d) The mixture spectrogram overlaid by the ideal binary mask. The recordings correspond to the left ear microphone.

where $s_T(t)$ is the reverberant target signal resynthesized through an all-one mask, $s_E(t)$ is the reverberant target signal resynthesized through the estimated binary mask, and $n_E(t)$ is the noise signal resynthesized through the same mask. While the output SNR measures the level of noise that remains in the reconstructed signal, the RSR measures the percentage of target energy loss. The results, averaged across the ten input signals, show SNR improvements in the range of 8-11 dB while preserving much of the target energy (~70-90%) for input SNR levels greater than or equal to 0 dB. Observe that the system performance degrades at lower SNR values because of increased overlap between target and interference. The RSR may be improved by imposing a higher threshold on the output-to-input attenuation level at the expense of increasing the residual noise in the output signal.

Table 1: SNR evaluation for a one-source interference (Scene 1)

Input SNR	-5 dB	0 dB	5 dB	10 dB
Output SNR (dB)	6.36	11.55	15.87	19.69
RSR (%)	59	74	84	91

Table 2: SNR evaluation for a two-speaker interference (Scene 2)

Input SNR	-5 dB	0 dB	5 dB	10 dB
Output SNR (dB)	4.82	10.18	14.68	18.54
RSR (%)	58	73	83	90

Table 3: SNR evaluation for a four-speaker interference (Scene 3)

Input SNR	-5 dB	0 dB	5 dB	10 dB
Output SNR (dB)	3.41	8.94	13.68	17.79
RSR (%)	52	66	79	89

Table 4 shows the performance of our system for six reverberation times between 0.0 s (anechoic) and 0.5 s (e.g., large living rooms and classrooms) which are obtained by simulating room impulse responses with different room absorption characteristics. Results are reported for Scene 1 and 0 dB input SNR. For each room configuration, the filter in the TCM module is adapted using 10 s of white noise simulated at target location as mentioned earlier. Overall, the system performance degrades by 8 dB output SNR when T_{60} is 0.2 s compared to the anechoic case while preserving the same retained speech ratio. This is partly due to the spectral smearing of individual sources as the reverberation time increases which results in increased overlap between target and interference. However, note that the RSR is above 70% across all conditions.

Table 4: SNR evaluation at different reverberation levels for a one source interference and 0 dB input SNR

	Output SNR (dB)	RSR (%)
T_{60}=0.0 s	18.74	70
T_{60}=0.1 s	13.14	73
T_{60}=0.2 s	10.89	74
T_{60}=0.3 s	11.55	74
T_{60}=0.4 s	11.49	75
T_{60}=0.5 s	10.99	74

We compare the performance of our algorithm with the standard delay-and-sum beamformer which is computationally simple and requires no knowledge about the interfering sources. As discussed in the introduction, while fixed beamformers are computationally simple and require only the target direction, they require a large number of microphones to obtain a good resolution. For our two-microphone configuration, the delay-and-sum beamformer produces only an average of 1.2 dB SNR gain across all three conditions.

To compare our model with adaptive beamforming techniques, we have implemented the two-stage adaptive filtering strategy described in Van Compernolle (1990) that improves the classic Griffith-Jim model under reverberation. The first stage is identical to our target cancellation module and is used to obtain a good noise reference. The second stage uses another adaptive filter to model the difference between the noise reference and the noise portion in the primary microphone. Here, training for the second filter is done independently for each noise condition (Scene 1 - Scene 3) in the absence of target signal using 10 s white noise sequences presented at each location in the tested configuration. The length of the filter is the same as the one used in the TCM (375 ms). Note that this approach requires adaptation for any change in both target source location as well as any interfering source location. As expected, the adaptive beamformer is optimal for canceling out one interfering source and hence gives an SNR gain of 13.61 dB in the Scene 1 condition. However, the second adaptive filter is not able to adapt to the noise configuration when multiple interferences are active since each source has a different differential path between the microphones. The adaptive beamformer thus produces an SNR gain of 3.63 dB in the Scene 2 condition and only 2.74 dB in the Scene 3 condition. The advantage for both the fixed beamformer as well as the adaptive one is that target signal distortions are minimal in the output when the filters are calibrated. By comparison, our system introduces some target energy loss. However, note that in the Scene 3 condition our system produces an SNR gain of 8 dB while losing less than 30% energy in the target signal for input SNR levels greater than 0 dB.

Given our computational objective of estimating the ideal binary mask, we also employ an SNR evaluation that uses the signal reconstructed from the ideal binary mask as ground truth (see Hu and Wang, 2004):

$$SNR_{IBM} = 10\log 10 \frac{\sum_t s_{IBM}^2(t)}{\sum_t (s_{IBM}(t) - s_E(t))^2}, \tag{6}$$

where $s_{IBM}(t)$ represents the target signal reconstructed using the ideal binary mask and $s_E(t)$ is the estimated target reconstructed from the binary mask produced by our model. Table 5 provides a comparison between our proposed system and the adaptive beamformer approach described above using this SNR measure. In order to extend the evaluation to the adaptive beamformer, the waveform at the beamformer output needs to be converted into a binary mask representation. Assuming target energy and noise energy are uncorrelated in individual T-F units, we can construct a binary mask as follows. For each T-F unit, if the energy ratio between the beamformer output and the input mixture is greater than 0.5 we label the unit as 1, otherwise we label the unit as 0. The signal resynthesized by applying this mask to the output waveform is used in Eq. (6) as the estimated target. As seen in the table, our system provides consistent improvements over the adaptive beamformer in low input SNR scenarios with multiple interferences (Scene 2 and Scene 3).

Table 5: Comparison with adaptive beamforming in terms of SNR with adaptive beamforming

	Input SNR	-5 dB	0 dB	5 dB	10 dB
Scene 1	Adaptive Beamformer	6.43	8.83	11.34	13.90
	Proposed System	3.72	6.47	8.92	11.70
Scene 2	Adaptive Beamformer	-0.40	4.22	8.47	12.38
	Proposed System	2.94	5.85	8.53	11.33
Scene 3	Adaptive Beamformer	-1.51	3.18	7.56	11.75
	Proposed System	2.14	4.88	7.58	10.69

A combination of target cancellation using a first-order differential beamformer and a spectral subtraction technique has been proposed previously by Álvarez et al. (2002). Since the first stage of our system produces a noise estimate, alternatively we can combine our adaptive filtering stage with spectral subtraction to enhance the reverberant target signal. However, as we will show in the following subsection, the computation of the binary mask improves front-end robustness compared to spectral subtraction in ASR applications.

B. ASR evaluation

We also evaluate the performance of our system as a front-end to a robust ASR system. The task domain is speaker independent recognition of connected digits. Thirteen (the numbers 1-9, a silence, very short pause between words, zero and oh) word-level models are trained using an HMM toolkit, HTK (Young et al., 2000). All except the short pause model have 8 emitting states. The short pause model has a single emitting state, tied to the middle state of the silence model. The output distribution in each state is modeled as a mixture of 10 Gaussians. The HMM architecture is the same as the one used in Cooke et al. (2001). The grammar for this task allows for one or more repetitions of digits and all digits are equally probable. Training is performed using the 4235 clean signals from the male speaker dataset in the TIDigits database (Leonard, 1984) downsampled to 16 kHz to be consistent with our model. Testing is performed on a subset of the testing set containing 229 utterances from 3 speakers which is similar to the test set used in Palomäki et al. (2004). The test speakers are different from the speakers in the training set. The test signals are convolved with the corresponding left and right ear target impulse responses and noise is added as described above to simulate the three conditions Scene 1-Scene 3.

We have trained the above HMMs with clean utterances from the training data using feature vectors consisting of the 13 mel-frequency cepstral coefficients (MFCC) together with their first and second order temporal derivatives. MFCCs are used as feature vectors as they are most commonly used in state-of-the-art recognizers (Rabiner and Juang, 1993). Cepstral mean normalization (CMN) is applied to the cepstral features in order to improve the robustness of the system under reverberant conditions (Shire, 2000). Frames are extracted using 20 ms windows with 10 ms overlap. A first-order preemphasis coefficient of 0.97 is applied to the signal. The recognition accuracy using clean test utterances is 99%. Using the reverberated test utterances, performance degrades to 94% accuracy.

The CMN applied on the MFCC features provides a relatively robust front-end for our task domain under the moderate reverberant conditions considered here. Hence, a reasonable approach is to remove the noise component from our acoustic mixture in the front-end processor and to feed an estimate of the reverberant target to the MFCC-based ASR. Although subjective listening tests have shown that the signal reconstructed from the ideal binary mask is highly intelligible (Roman et al., 2003; Chang 2004; Brungart et al., 2005), the extraction of MFCC features from a signal reconstructed using such a mask is distorted due to the mismatch arising from the T-F units labeled 0, which smears the entire cepstrum via the cepstral transform (Cooke et al., 2001). A similar problem occurs when the second stage of our model is replaced by spectral subtraction since spectral subtraction performs poorly in the T-F regions dominated by interference where oversubtraction or undersubtraction occurs. One way to handle this problem is by estimating the original target spectral values in the T-F units labeled 0 using a prior speech model. This approach has been suggested by Raj et al. (2004) in the context of additive noise. In this approach, a noisy spectral vector Y at a particular frame is partitioned in its reliable Y_r and its unreliable Y_u components. The task is to reconstruct the underlying true spectral vector X. Assuming that the reliable features Y_r are approximating well the true ones X_r, a Bayesian decision is then employed to estimate the remaining X_u given only the reliable component. Hence, this approach works seamlessly with the T-F binary mask that our speech segregation system produces. Here, the reliable features are the T-F units labeled 1 in the mask while the unreliable features are the ones labeled 0. We train the prior speech model on clean data to avoid obtaining a prior for each deployment condition which is desirable for robust speech recognition.

The speech prior is modeled empirically as a mixture of Gaussians and trained with the same clean utterances used in ASR training:

$$p(X) = \sum_{k=1}^{M} p(k) p(X \mid k) \tag{7}$$

where M=1024 is the number of mixtures, k is the mixture index, $p(k)$ is the mixture weight and $p(X \mid k) = N(X; \mu_k, \Sigma_k)$.

Previous studies (Cooke et al., 2001; Raj et al., 2004) have shown that a good estimate of X_u is its expected value conditioned on X_r:

$$E_{X_u \mid X_r, 0 \le X_u \le Y_u}(X_u) = \sum_{k=1}^{M} p(k \mid X_r, 0 \le X_u \le Y_u) \int_0^{Y_u} X_u p(X_u \mid k, 0 \le X_u \le Y_u) dX_u \tag{8}$$

where $p(k| X_r)$ is the *a posteriori* probability of the k'th Gaussian given the reliable data and the integral denotes the expectation $\overline{X}_{u,k}$ corresponding to the k'th mixture. Note that under the additive noise condition, the unreliable parts may be constrained as $0 \le X_u \le Y_u$ (Cooke et al., 2001). In our implementation, we have assumed that the prior can be modeled using a mixture of Gaussians with diagonal covariance. Theoretically, this is a good approximation if an adequate number of mixtures are used. Additionally, our empirical evaluations have shown that for the case of M=1024 this approximation results in an insignificant degradation in recognition performance while the computational cost is greatly reduced. Hence, the expected value can now be computed as:

$$\tilde{X}_u = \begin{cases} \mu_{u,k} \ , \ 0 \le \mu_{u,k} \le Y_u \\ Y_u \ , \ \mu_{u,k} > Y_u \\ 0 \ , \ \mu_{u,k} < 0 \end{cases} \tag{9}$$

The *a posteriori* probability of the k'th mixture given the reliable data is estimated using the Bayesian rule from the simplified marginal distribution $p(X_r \mid k) = N(X_r; \mu_{r,k}, \sigma_{r,k})$ obtained from $p(X \mid k)$ without utilizing any bounds on X_u. While this simplification results in a small decrease in accuracy, it results in a substantially faster computation of the marginal.

Speech recognition results for the three conditions Scene 1, Scene 2 and Scene 3 are reported separately in Fig. 4, Fig. 5 and Fig. 6 at five SNR levels: –5 dB, 0 dB, 5 dB, 10 dB and 20 dB. Results are obtained using the same MFCC-based ASR as the back-end for the following approaches: fixed beamforming, adaptive beamforming, target cancellation through adaptive filtering followed by spectral subtraction, our proposed front-end ASR using the estimated mask and finally our proposed front-end ASR using the ideal binary mask. The baseline results correspond to the unprocessed left ear signal. Observe that our system achieves large improvements over the baseline performance across all conditions. Additionally, the excellent results reported for the ideal binary mask highlights the potential performance that can be obtained using this approach. Note that the ASR performance depends on the interference type and we obtain the best accuracy score in the two-speaker interference condition. As seen also in the SNR evaluation, the adaptive beamformer outperforms all the other algorithms in the case of a single interference (Scene 1). However, as the number of interferences increases, the performance of the adaptive beamformer degrades rapidly and approaches the performance of the fixed beamformer in the Scene 3 condition. As described in the previous subsection, we can combine our adaptive filtering stage with spectral subtraction to cancel the interference. As illustrated by the recognition results in Fig. 5 and Fig. 6, this approach outperforms the adaptive beamformer in the case of multiple concurrent interferences. While spectral subtraction improves the SNR gain in target-dominant T-F units, it does not produce a good target signal estimate in noise-dominant regions. Note that our front-end ASR employs a better estimation of the spectrum in the unreliable T-F units and therefore results in large improvements over the spectral subtraction method.

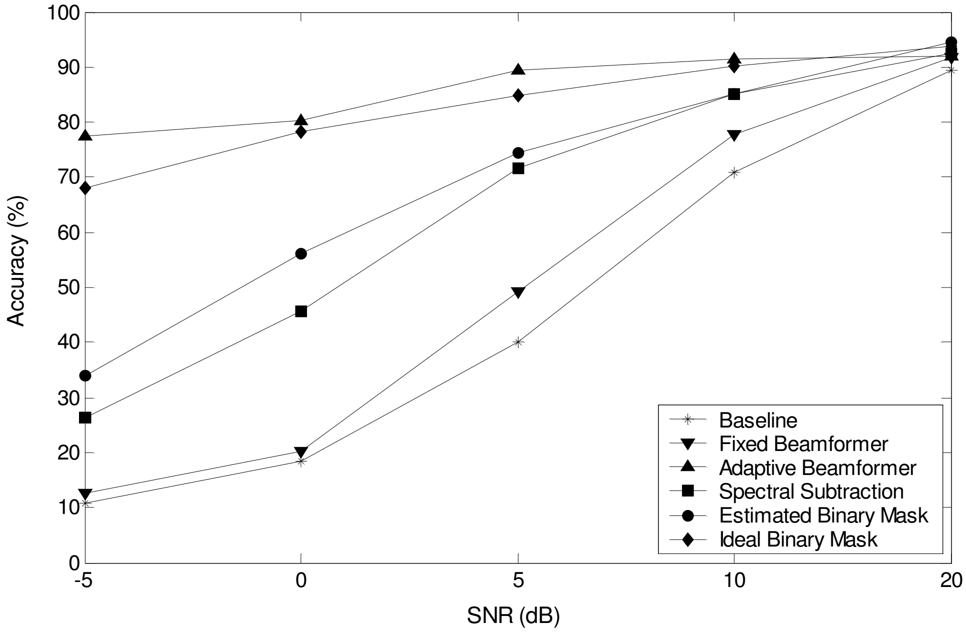

Figure 4. Recognition performance for Scene 1 at different SNR values for the reverberant mixture (*), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), our front end ASR using the estimated binary mask (●), our front-end ASR using the ideal binary mask (♦).

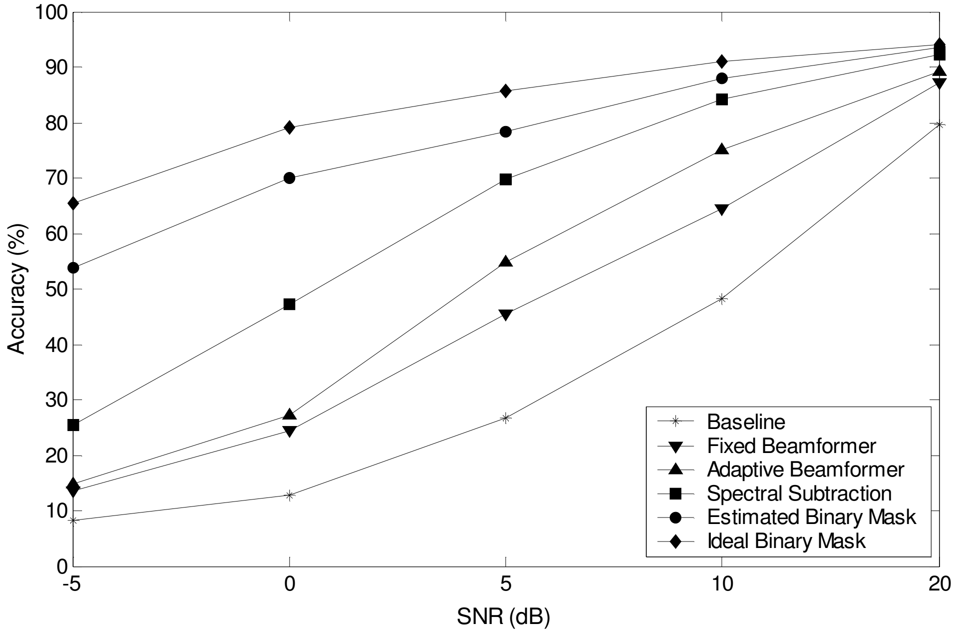

Figure 5. Recognition performance for Scene 2 at different SNR values for the reverberant mixture (*), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), our front end ASR using the estimated binary mask (●), our front-end ASR using the ideal binary mask (♦).

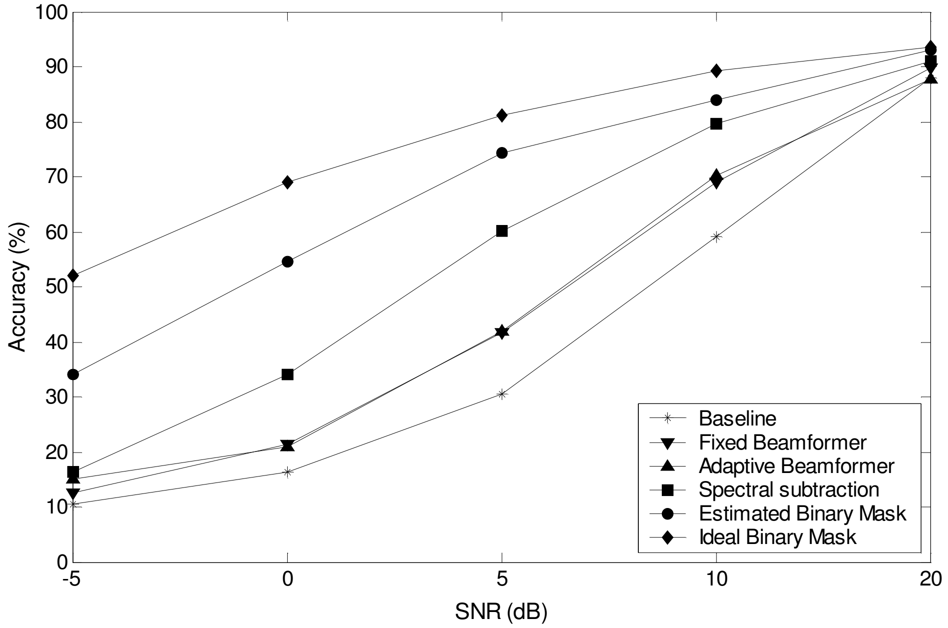

Figure 6. Recognition performance for Scene 3 at different SNR values for the reverberant mixture (*), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), our front end ASR using the estimated binary mask (●), our front-end ASR using the ideal binary mask (♦).

We compare our system with the binaural system proposed by Palomäki et al. (2004) which was shown to produce substantial recognition improvements on the same digit recognition task as used here. Their system combines binaural localization with precedence effect processing in order to detect reliable spectral regions that are not contaminated by interfering noise or echoes. Recognition is then performed in the log spectral domain by employing the missing data ASR system proposed by Cooke et al. (2001). This recognizer takes as input a binary mask that identifies the reliable data in the mixture spectrogram and uses this to compute the state output probabilities for each observed vector based only on its reliable parts. In order to account for the reverberant environment, spectral energy normalization is employed. While our system can handle a variety of interfering sources, the binaural system of Palomaki et al. was developed for only one-interference scenarios. Table 6 compares the two systems for the case of one interfering source of rock music, which was used in Palomäki et al. The recognition results for the Palomäki et al. system are the ones reported by the authors while the results for our system have been produced using the following configuration setup. Listener is located in the middle of the room while target and interfering sources are located at 20° and -20° respectively. T_{60} is 0.3 s and the input SNR is fixed before the binaural presentation of the signals at three SNR levels: 0 dB, 10 dB and 20 dB. Note that we obtain a marked improvement over the system of Palomäki et al. (2004), in the low SNR conditions. By utilizing ITD and IID information only during acoustic onsets, the mask obtained by their system has a limited number of reliable units. This limits the amount of information available to the missing data recognizer for the decoding (Srinivasan et al., 2004). In our system, on the other hand, a novel encoding of the target source location leads to the recovery of more target dominant regions and this results in a more robust front-end for ASR.

Table 6: Comparison with the Palomäki et al. system in terms of speech recognition accuracy (%)

Input SNR	0 dB	10 dB	20 dB
Baseline (MFCC+CMN)	13.04	43.01	81.85
Palomäki et al. system	32.7	78.8	91.9
Proposed system	47.58	81.59	91.80

We further compare our system with the negative beamforming approach proposed by Álvarez et al. (2002) and the results are reported in Table 7. In order to compare with this approach we simulate the input for a two-microphone array with 5 cm intermicrophone distance using the image reverberation model (Allen and Berkley, 1979). We use the same room configuration, the same interfering signals and the same spatial configuration as in the Scene 3 condition described previously. The system proposed by Álvarez et al. uses a first-order differential beamformer to cancel the direct path of the target signal. Since target is fixed at 0°, the adaptation parameter in the differential beamformer is fixed to 0.5 across all frequencies (see Álvarez et al., 2002). The output of the differential beamformer contains both the reverberant path of the target signal as well as an estimate of the additional interfering sources. An additional frequency-equalizing curve is applied on this output since the amount of attenuation performed by this beamformer varies with the frequency of the signal as well as its location. This equalizing-curve is trained using white noise at the corresponding interfering locations. The estimated noise spectrum is finally subtracted from the spectrum of one of the two microphone mixtures (the left one) and the results are fed to the MFCC-based ASR. Our system is trained on the new configuration to obtain the TCM adaptive filter using the parameters described in Section II. The T-F mask produced by our system is then used to reconstruct the spectrogram using the prior speech model. As shown in Table 7, our system significantly outperforms the system of Álvarez et al. (2002) across a range of SNRs.

Table 7: Comparison with the Álvarez et al. system in terms of speech recognition accuracy (%)

Input SNR	0 dB	10 dB	20 dB
Baseline (MFCC+CMN)	11.69	40.99	82.80
Álvarez et al. system	24.14	51.61	73.39
Proposed system	31.59	75.00	91.94

IV. DISCUSSION

In natural settings, reverberation alters many of the acoustical properties of a sound source reaching our ears, including smearing of the binaural cues due to the presence of multiple reflections. This is especially detrimental when multiple sound sources are present in the acoustic scene since the acoustic cues are now required to distinguish between the competing sources. Location based algorithms that rely on the anechoic assumption of time delayed and attenuated mixtures are therefore prone to failure in reverberant scenarios. An adaptive filter can therefore be used to better characterize the target location in a reverberant room. We have presented here a novel two-microphone sound segregation system that performs well under such realistic conditions. Our approach is based on target cancellation through adaptive filtering followed by an analysis of the output-to-input attenuation level in individual T-F units. The output of the system is an estimate of an ideal binary mask which labels the T-F components of the acoustic scene dominated by the target sound.

Classical two-microphone noise cancellation strategies process the input using linear adaptive filters and while being optimal in the one-interference condition, they are unable to cope with multiple interferences. By using a non-linear strategy in the second stage, our system is able to cancel an arbitrary number of interferences using only two microphones. As shown in our SNR evaluation, the system is able to outperform existing beamforming techniques across a range of input SNRs. Note that while our processing produces some target signal distortion, we preserve most of the target energy (>70%) at input SNRs greater than 0 dB. Further, informal listening tests show that the reconstructed signals are highly intelligible. The balance between noise cancellation and target distortion can be controlled in our system by varying the output-to-input attenuation threshold. In high SNR conditions, for example, a more relaxed threshold will ensure less target distortion at the expense of some background noise.

Since the first stage of our system provides a noise estimate, an alternative non-linear strategy for the second stage is spectral subtraction. A combination of target cancellation through differential beamforming and spectral subtraction has been proposed previously by Alvarez et al. (2002). Informal listening tests have shown that the signals obtained using spectral subtraction in the second stage of your system have a similar quality to the ones resynthesized from the estimated binary masks. An SNR evaluation using the reverberant target as signal shows a slight improvement using the spectral subtraction method. However, as seen in the ASR evaluation, the binary masks provide a complimentary advantage when coupled with missing data techniques and can provide sizeable ASR improvements. Spectral subtraction, however, can also be used in combination with our binary mask estimation. We observe additional improvements (an absolute word error rate reduction of 3%-5%) when using spectral subtraction to “clean” the reliable regions prior to spectrogram reconstruction.

In terms of application to real-world scenarios our adaptive filtering strategy has several drawbacks. First, the adaptation of the inverse filter requires data on the order of a few seconds and thus any fast change in target location (e.g. walking) will have an adverse impact on the system. Second, the system needs to identify signal intervals that contain no interference to allow for the filter to adapt to a new target position**.** On the other hand, note that our system requires training only with respect to target location and is therefore insensitive to changes in the locations of interfering sources, unlike adaptive beamforming whose training is conditioned on the positions of all sound sources in the environment.

We use the approach proposed by Raj et al. (2004) to reconstruct the clean target signal in the unreliable T-F units. This allows for our system to be utilized as a front-end to a standard speech recognition system operating using cepstral features. In a systematic comparison, our system shows substantial performance gains over baseline and significant improvements over related approaches. Note that our prior and ASR models are trained on clean speech and hence our algorithm is applicable when recognition in changing reverberant environments is desired. However, if samples of reverberant target are available or dereverberation techniques become applicable we can utilize these to further improve the ASR performance.

ACKNOWLEDGMENTS

This research was supported in part by an AFOSR grant (FA9550-04-1-0117), an NSF grant (IIS-0081058) and an AFRL grant via Veridian. We thank B. Raj for helpful discussions. A preliminary version of this work is included in the Proceedings of 2004 ICASSP.

REFERENCES

J. B. Allen and D. A. Berkley (**1979**). “Image method for efficiently simulating small-room acoustics,” J. Acoust. Soc. Am., vol. 65, pp. 943-950.

A. Álvarez, P. Gómez, V. Nieto, R. Martínez and V. Rodellar (**2002**). “Speech enhancement and source separation supported by negative beamforming filtering,” Proc. International Conference on Signal Processing, pp. 342-345.

R. Balan, A. Jourjine and J. Rosca (**1999**). “AR processes and sources can be reconstructed from degenerate mixtures,” Proc. 1st International Workshop on Independent Component Analysis and Signal Separation, pp. 467-472.

A. K. Barros, T. Rutkowski, F. Itakura and N. Ohnishi (**2002**). “Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets,” IEEE Trans. Neural Net., vol. 13, pp. 888-893.

M. Brandstein and D. Ward, Eds. (**2001**). *Microphone Arrays: Signal Processing Techniques and Application*, Berlin: Springer.

A. S. Bregman (**1990**). *Auditory Scene Analysis, Cambridge*, MA: MIT press.

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

D. Brungart, P. Chang, B. Simpson and D. L. Wang (**2005**). “Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask,” submitted for journal publication.

M. D. Burkhard and R. M. Sachs (**1975**). “Anthropometric manikin for acoustic research,” J. Acoust. Soc. Am., vol. 58, pp. 214-222.

P. Chang (**2004**). “Exploration of behavioral, physiological, and computational approaches to auditory scene analysis,” M.S. thesis, The Ohio State University. Available at http://www.cse.ohio-state.edu/pnl/theses/Chang MSThesis04.pdf

M. P. Cooke, P. Green, L. Josifovski and A. Vizinho (**2001**). “Robust automatic speech recognition with missing and unreliable acoustic data,” Speech Comm., vol. 34, pp. 267-285.

P. Divenyi, Ed. (**2005**). Speech Separation by Humans and Machines, Norwell MA: Kluwer Academic.

Y. Ephraim and H. L. Trees (**1995**). "A signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Proc., vol. 3, pp. 251-266.

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

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

L. J., Griffiths and C. W. Jim (**1982**). "An alternative approach to linearly constrained adaptive beamforming," IEEE Trans. Antennas Propag., vol. 30, pp. 27-34.

S. Haykin (**2002**). *Adaptive Filter Theory*, 4th ed., Upper Saddle River, New Jersey: Prentice Hall.

A. Hyväarinen, J. Karhunen and E. Oja (**2001**). *Independent Component Analysis*, New York: Wiley.

G. Hu and D. L. Wang (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Net., vol. 15, pp. 1135-1150.

R. G. Leonard (**1984**). "A database for speaker-independent digit recognition," Proc. ICASSP, pp. 111-114.

C. Liu, B. C. Wheeler, W. D., Jr., O'Brien, C. R. Lansing, R. C. Bilger, D. L. Jones and A. S. Feng (**2001**). "A two microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," J. Acoust. Soc. Am., vol. 110, pp. 3218-3230.

H. Y. Luo and P. N. Denbigh (**1994**). "A speech separation system that is robust to reverberation," Proc. International Symposium on Speech, Image Processing, and Neural Networks, pp. 339-342.

N. Ma, M. Bouchard and R. Goubran (**2004**). "Perceptual Kalman filtering for speech enhancement in colored noise", Proc. ICASSP, vol. 1, pp.717-720.

R. Martin (**2001**). "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Proc., vol. 9, pp. 504-512.

K. J. Palomäki, G. J. Brown and D. L. Wang (**2004**). "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," Speech Comm., vol. 43, pp. 361-378.

L. R. Rabiner and B. H. Juang (**1993**). *Fundamentals of Speech Recognition*, 2nd ed., Englewood Cliffs, NJ: Prentice-Hall.

B. Raj, M. L. Seltzer, R. M. Stern (**2004**). “Reconstruction of missing features for robust speech recognition,” Speech Comm., vol. 43, pp. 275-296.

N. Roman and D. L. Wang (**2004**). "Binaural sound segregation for multisource reverberant environments," Proc. ICASSP, vol.2, pp. 373-376

N. Roman, D. L. Wang and G. J. Brown (**2003**). “Speech segregation based on sound localization,” J. Acoust. Soc. Am., vol. 114, pp. 2236-2252.

A. Shamsoddini and P. N. Denbigh (**2001**). “A sound segregation algorithm for reverberant conditions,” Speech Comm., vol. 33, pp. 179-196.

M. L. Shire (**2000**). “Discriminant training of front-end and acoustic modeling stages to heterogeneous acoustic environments for multi-stream automatic speech recognition,” Ph.D. dissertation, University of California, Berkeley.

S. Srinivasan, N. Roman and D. L. Wang (**2004**). "On binary and ratio time-frequency masks for robust speech recognition,” Proc. ICSLP, pp. 2541-2544.

D. Van Compernolle (**1990**). “Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings,” Proc. ICASSP, pp. 833-836.

D. L. Wang (**2005**). “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, P. Divenyi, ed., Norwell MA: Kluwer Academic, pp. 181-197.

D. B. Ward, R. A. Kennedy and R. C. Williamson (**2001**). “Constant directivity beamforming,” in *Microphone Arrays: Signal Processing Techniques and Application*, M. Brandstein and D. Ward, eds., Berlin: Springer, pp. 3-17.

M. Weiss (**1987**). “Use of an adaptive noise canceller as an input preprocessor for a hearing aid,” J. Rehab. Res. Devel., vol. 24, pp. 93-102.

B. Widrow and S. D. Stearns (**1985**). *Adaptive Signal Processing*, New Jersey: Prentice-Hall.

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

M. Zibulevsky, B. A. Pearlmutter, P. Bofill and P. Kisilev (**2001**). "Blind source separation by sparse decomposition", in Independent Component Analysis: Principles and Practice, Roberts, S. J., and Everson, R.M., Eds., Cambridge University Press.