14

Binaural Speech Segregation

Nicoleta Roman[1] and DeLiang Wang[2]

[1] Ohio State University at Lima, Lima, USA
[2] Ohio State University, Columbus, USA

It is relatively easy for a human listener to attend to a particular speaker at a cocktail party in the presence of other speakers, music and environmental sounds. To perform this task, the human listener needs to separate the target speech from a mixture of multiple concurrent sources reflected by various surfaces. This process is referred to as *auditory scene analysis*. While humans excel at this task using only two ears, machine separation based on two-microphone recordings has proven to be extremely challenging. By incorporating the mechanisms underlying the perception of sound by human listeners, *computational auditory scene analysis* (CASA) offers a new approach to sound segregation. Binaural hearing – hearing with two ears – employs the difference in sound source locations to improve sound segregation. In this chapter, we describe the principles of binaural processing and review the state-of-the-art in binaural CASA, particularly for speech segregation.

14.1 Introduction

Human listeners are able to effectively process the multitude of acoustic events that surrounds them at all times. Each acoustic source generates a vibration of the medium (air) and our hearing is confronted by the superposition of all vibrations impinging on our eardrums. As Helmholtz noted in 1863, the final waveform is “complicated beyond conception” [26]. Nonetheless, at a cocktail party, we are able to attend to and understand a particular talker. This perceptual ability is known as the “cocktail-party effect” – a term introduced by Cherry in 1953 [15]. Cherry’s original experiments have triggered research in widely different areas including speech perception in noise, selective attention, neural modeling, speech enhancement and source separation. Of special interest is a machine solution to the problem of sound separation in realistic environments, which is essential to many important applications including automatic speech and speaker recognition, hearing aid design and audio information retrieval. The field of automatic speech recognition (ASR), for example, has seen much progress in recent years. However, the performance of current recognition systems degrades rapidly in the presence of noise and reverberation and the degradation is much faster compared to human performance in similar conditions [30, 33].

The sound separation problem has been investigated in the signal processing field for many years for both one-microphone recordings and multi-microphone ones (for recent reviews see [8,19]). One-microphone speech enhancement techniques include spectral subtraction [40], Kalman filtering [38], subspace analysis [20] and autoregressive modeling [5]. While requiring only one sensor benefits many applications, these algorithms make strong assumptions about interference and thus have difficulty in dealing with general acoustic mixtures. Microphone array algorithms include beamforming [8] and blind source separation (BSS) through independent component analysis (ICA) [31]. To separate multiple sound sources, beamforming takes advantage of their different directions of arrival while ICA relies on their statistical independence. The main drawback of these approaches is that they generally require the number of microphones equal or exceed the number of sound sources. In the case of two-microphone recordings typically only one wideband interfering source can be canceled out by steering a null towards its location. To address this problem, it has been proposed in [35] a subband adaptive beamformer which steers independent nulls in each time-frequency (T-F) unit to suppress the strongest interference. Another approach to this underdetermined problem – more sources than sensors – is sparse signal representation based ICA [64]. The performance of these approaches is still limited in realistic multi-source reverberant conditions.

Since the natural solution provided by human hearing is robust to noise and reverberation, one can expect that a solution to the sound separation problem can be devised using up to two microphones. While human listeners can separate speech monaurally, binaural hearing adds to this ability when sources are spatially separated [10]. A coherent theory on the human ability to segregate signals from noisy mixtures was presented by Bregman in 1990 [9]. He argues that humans perform an auditory scene analysis (ASA) of the acoustic input in order to form perceptual representations of individual sources called streams. ASA takes place in two stages: the first stage decomposes the input into a collection of sensory elements while the second stage selectively groups the elements into streams that correspond to individual sound sources. According to Bregman, stream segregation is guided by a variety of grouping cues including proximity in frequency and time, pitch, onset/offset, and spatial location. The 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 binaural separation [12, 28, 60]. Mirroring the ASA processing described above, CASA systems generally employ two stages: segmentation (analysis) and grouping (synthesis).

In segmentation, the acoustic input is decomposed into sensory segments, or contiguous T-F regions, each of which mainly originates from a single source. In grouping, segments that are likely to come from the same source are put together. Most monaural segregation algorithms rely on pitch as the main grouping cue and therefore can operate only on voiced speech or other periodic sounds (e.g., [27]; see also [29] for an exception). On the other hand, binaural algorithms employ location cues which are independent of signal content and thus can be used to track both voiced and unvoiced speech. Compared with signal processing techniques, CASA systems make relatively few assumptions about the acoustic properties of the interference and the environment.

CASA systems typically employ T-F masking to segregate target signal from mixture signal [11,59,60]. Specifically, T-F units in the acoustic mixture are selectively weighted in order to enhance the desired signal. The weights can be binary or real [53]. T-F binary masking is motivated by the masking phenomenon in human audition, in which a weaker signal is masked by a stronger one in the same critical band [41]. Subsequently, the computational goal of a CASA system has been argued to be an ideal T-F binary mask, which selects the target if it is stronger than the interference in a local T-F unit [47, 49, 58]. Speech extracted from such masks has been shown to be highly intelligible in multi-source mixtures [14, 49], as well as to produce substantial improvements in robust speech recognition [17,49]. Following the ASA account, the binary mask is estimated in a CASA system by grouping T-F units using various perceptual cues. This binary masking framework has recently become popular in the underdetermined BSS field as well, as it has been observed that different speech signals can be approximately orthogonal in a high-resolution T-F representation [3,44,62]. In [44] for example, ICA is combined with T-F binary masking to iteratively extract speech signals from underdetermined anechoic mixtures.

The binaural cues used by the auditory system for source localization are interaural time difference (ITD) and interaural intensity difference (IID) between the two ears [6]. While filtering produced by head, torso and external ear introduce only a weak frequency dependency for ITD [39], IID varies widely across frequencies ranging from 0.5-1 dB at low frequencies to as much as 30 dB at high frequencies. Consequently, while ITD is the main localization cue employed by the auditory system at lower frequencies (<1.5 kHz), both binaural cues are used at higher frequencies. A series of psychoacoustically inspired binaural processors have shown that these location cues can be used to substantially enhance target speech in binaural mixtures [7,37] [61]. Recently, binaural CASA systems have employed supervised learning in the ITD-IID feature space to optimally extract the target signal [13, 49, 53]. The main observation is that, in a given T-F unit, there exists a systematic relationship between the *a priori* local SNR and the deviation of ITD/IID features [49]. Moreover, in the case of multiple concurrent sources, there exists a characteristic clustering in the ITD/IID space which can be used to estimate the ideal T-F binary mask. Systematic evaluations have shown that systems developed based on these observations perform very well under multi-source anechoic conditions [49, 53].

Reverberation presents an additional challenge to a binaural system as it introduces potentially an infinite number of additional sources due to reflections from hard surfaces. This smears considerably the ITD/IID statistics. As a result, the performance of the above location-based segregation systems degrades as reverberation level increases. Inspired by psychoacoustical studies, many systems use a model of precedence effect prior to binaural processing to emphasize the cues in the direct wavefront over the cues in the later reflections [13,43]. Alternatively, we have proposed to replace the anechoic modeling of ITD/IID with an adaptive filter to better characterize the target location in reverberant conditions [50]. The system in [50] performs target cancellation through adaptive filtering followed by an analysis of the output-to-input attenuation level to estimate the ideal binary mask. A systematic evaluation shows that the system results in large SNR gains and it outperforms standard two-microphone beamforming algorithms as well as a recent binaural processor.

The rest of the chapter is organized as follows. The next section describes T-F masks for CASA systems. Sec. 14.3 describes a binaural system for multi-source anechoic conditions. Sec. 14.4 describes a binaural system for multi-source reverberant conditions. Sec. 14.5 gives evaluation data for the systems described in Sec. 14.3 and Sec. 14.4. The last section concludes the chapter.

14.2 T–F Masks for CASA

The first stage of a CASA system is usually a T-F analysis of the input signal using either a physiologically motivated filterbank that mimics cochlear filtering [16] or a short-time Fourier transform (STFT). In this paper, we use an STFT representation to illustrate the concepts of T-F masking and binaural processing (see also the next two sections) but a similar description can be made using an auditory filterbank. Given a T-F decomposition of the acoustic mixture, the target source can be recovered by applying independent weights to individual T-F units. This type of T-F masking can be viewed as a nonstationary Wiener filter. The authors in [21] have shown that the minimum mean-square error estimate of the target signal amplitude in a T-F unit is related to the *a priori* local SNR. Hence, we define an ideal ratio mask using the *a priori* energy ratio as follows:

$$R(\Omega,t) = \frac{\left|S\left(e^{j\Omega},t\right)\right|^2}{\left|S\left(e^{j\Omega},t\right)\right|^2 + \left|N\left(e^{j\Omega},t\right)\right|^2}, \tag{14.1}$$

where $S(e^{j\Omega},t)$ is the spectral value for the target signal and $N(e^{j\Omega},t)$ is the spectral value for the interference at frequency Ω and frame index t.

As described previously, a number of researchers have shown the potential of binary T-F masking in speech segregation. The upper limit for a CASA system that uses binary masking is an ideal binary mask, which selects the T-F units where the target energy is stronger than the interference energy. Formally, this ideal binary mask is defined as follows:

$$M_{\mathrm{IBM}}(\Omega, t) = \begin{cases} 1, & \text{if } \left|S\left(e^{j\Omega}, t\right)\right| > \left|N\left(e^{j\Omega}, t\right)\right|, \\ 0, & \text{otherwise.} \end{cases} \tag{14.2}$$

This is equivalent to applying a threshold of 0.5 on the energy ratio $R(\Omega, t)$. By selecting the T-F units where the target is stronger than the interference, this definition results in the optimal SNR gain among all possible binary masks because the SNR in each T-F unit is positive if the unit is retained and negative if the unit is discarded [27]. Although an ideal ratio mask will outperform an ideal binary mask [53], the estimation of an ideal ratio mask has turned out to be more sensitive to corruptions by noise and reverberation than estimating the ideal binary mask. This chapter will therefore focus on the estimation of an ideal binary mask.

An important application for CASA systems is to provide a robust front-end for ASR. Given a T-F mask (binary or ratio), the target signal can be reconstructed using the element-wise multiplication of the mask and the spectral energy of the mixture. While the signal obtained from a ratio mask can be used directly as input to a speech recognizer, conventional ASR systems are highly sensitive to the distortions introduced by binary masks. Cooke et al. [17] have proposed a missing-data approach to ASR which performs recognition using only the reliable (clean) components. Hence, a binary mask which labels the T-F units where target dominates interference is therefore an ideal front-end for this approach. A number of authors have shown that the ideal binary masks used as front-ends to a missing-data ASR provide impressive results even under very low SNR conditions [17, 49]. Alternatively, Raj et al. [46] have proposed a spectral reconstruction method for the T-F units dominated by noise to alleviate the mismatch introduced by binary masking. The reconstructed signal is then used as input to a conventional ASR system. While the missing-data ASR requires spectral features, a conventional ASR usually employs cepstral features which are known to be more effective than the spectral ones.

14.3 Anechoic Binaural Segregation

Under anechoic conditions, the signal emitted by an acoustic source arrives at the ear further away from the source at a later time and attenuated compared to the signal arriving at the ear closer to the emitting source. Inspired by psychoacoustical studies of sound localization, binaural sound separation systems have typically employed the binaural cues of ITD and IID for localization and further segregation of target source [6, 7, 49, 53]. Specifically, the filtering due to head, pinna and torso introduces at each frequency natural combinations of ITD and IID which are location dependent. When the target source dominates a particular frequency bin, the observed ITD and IID correspond to the target values. When an interfering source overlaps with the target one in the same frequency bin, the observed ITD and IID undergo systematic shifts as the energy ratio between the two sources changes [49]. This relationship is used to estimate the weights independently in each T-F unit in order to extract target signal from noisy mixture.

The presentation here is based on the binaural system proposed in [53]. The ITD and IID estimates are computed based on the spectral ratio at the left and right ears:

$$\left(\widehat{ITD}, \widehat{IID}\right)(\Omega, t) = \left[-\frac{1}{\Omega} A\left(\frac{X_{\mathrm{L}}\left(e^{j\Omega}, t\right)}{X_{\mathrm{R}}\left(e^{j\Omega}, t\right)}\right), \left|\frac{X_{\mathrm{L}}\left(e^{j\Omega}, t\right)}{X_{\mathrm{R}}\left(e^{j\Omega}, t\right)}\right|\right] \quad (14.3)$$

where $X_{\mathrm{L}}(e^{j\Omega}, t)$ and $X_{\mathrm{R}}(e^{j\Omega}, t)$ are the left and right ear spectral values of the mixture signal at frequency Ω and frame t and $A(re^{j\phi})$=ϕ, -$\pi < \phi < \pi$. The function A computes the phase angle, in radians, of a complex number with magnitude r and phase angle ϕ. The phase is ambiguous corresponding to integer multiples of 2π. We therefore consider ITD in the range $2\pi/\Omega$ centered at zero delay.

A corpus of 10 speech signals from the TIMIT[3] database [23] is used for training. Five sentences correspond to the target location set and the rest belong to the interference location set. Binaural signals are obtained by convolving monaural signals with measured head-related impulse responses (HRIRs) corresponding to the direction of sound incidence. The responses to multiple sources are added at each ear. The HRIR measurements consist of left/right responses of a KEMAR[4] manikin from a distance of 1.4 m in the horizontal plane, resulting in 128 point impulse responses at a sampling rate of 44.1 kHz [22].

Fig. 14.1 shows empirical results from the above corpus for a two-source configuration: target source in the median plane and interference at 30°. The T-F resolution is 512 discrete Fourier transform (DFT) coefficients extracted every 20 ms with a 10 ms overlap. ITD/IID and energy ratio estimates are computed every frame using the formulas in Eqs. 14.1 and 14.3. The scatter plot in Fig. 14.1(a) shows samples of $\widehat{ITD}(\Omega, t)$ and $R(\Omega, t)$ for a frequency bin at 1 kHz. Similarly, Fig. 14.1(b) shows the results that describe the variation of $\widehat{IID}(\Omega, t)$ and $R(\Omega, t)$ for a frequency bin at 3.4 kHz. Note that the scatter plots in Fig. 14.1 exhibit a systematic shift of the estimated ITD and IID with respect to R. Moreover, a location-based clustering is observed in the joint ITD-IID space as shown in Fig. 14.1(c). Each peak in the histogram corresponds to a distinct active source.

[3] The term *TIMIT* results from *Texas Instruments (TI)* and *Massachusetts Institute of Technology (MIT).*

[4] *KEMAR* abbreviates *Knowles electronic manikin for acoustic research.*

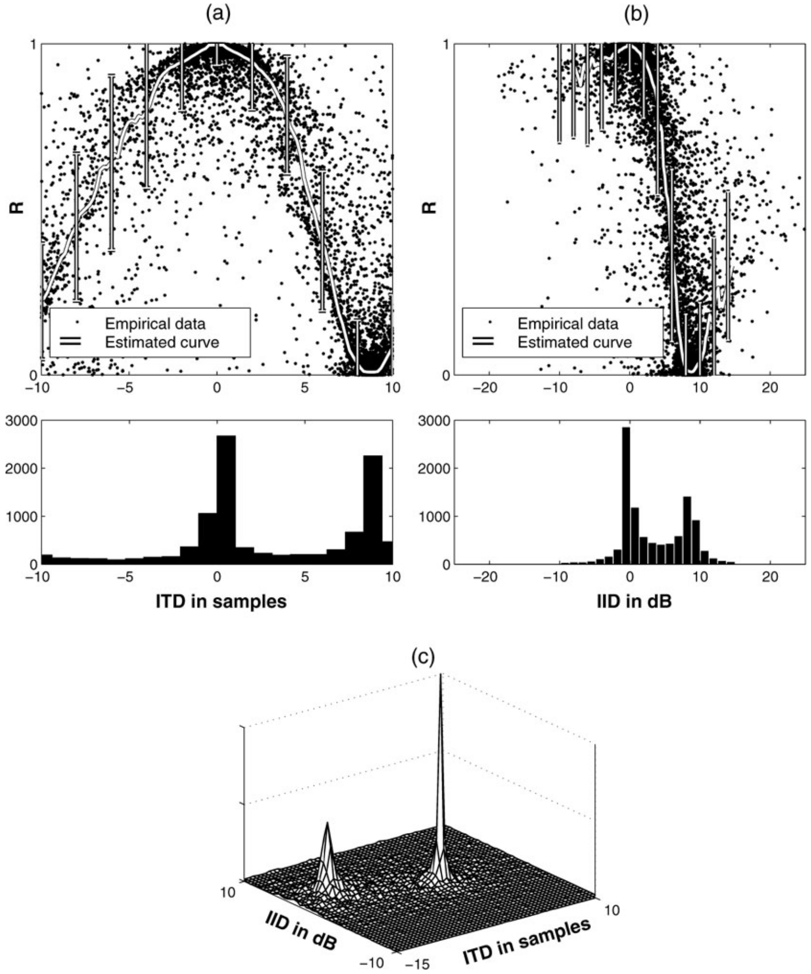

Fig. 14.1. Relationship between ITD/IID and the energy ratio R (from [53]). Statistics are obtained with target in the median plane and interference on the right side at 30°. (a) The top panel shows the scatter plot for the distribution of R with respect to ITD for a frequency bin at 1 kHz. The solid 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.

To estimate the ideal binary mask $M_{\mathrm{IBM}}(\Omega, t)$ we employ a non-parametric classification in the joint ITD-IID feature space. There are two hypotheses for the binary decision:

- H_1 – target is stronger or $R(\Omega, t) \geq 0.5$ and

- H_2 – interference is stronger or $R(\Omega, t) < 0.5$.

The estimated binary mask, $\widehat{M}_{\text{IBM}}(\Omega, t)$, is obtained using the maximum *a posteriori* (MAP) decision rule:

$$\widehat{M}_{\text{IBM}}(\Omega, t) = \begin{cases} 1, & \text{if} \quad p(H_1)p(x|H_1) > p(H_2)p(x|H_2), \\ 0, & \text{otherwise}, \end{cases} \tag{14.4}$$

where x corresponds to the ITD and IID estimates. 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 (see also [49]). Alternatively, the ITD/IID statistics can be used to derive ratio/soft masks [13,53]. In [53] for example, the empirical mean curves shown in Fig. 14.1 are used to estimate the energy ratio from the observed ITD/IID.

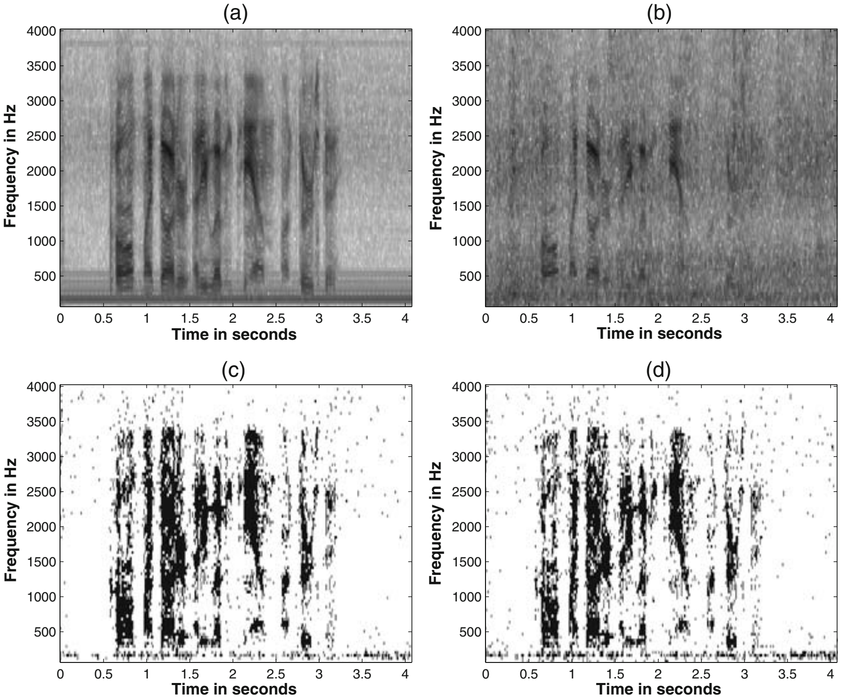

Fig. 14.2. Comparison between estimated and ideal T-F binary masks for a mixture of speech utterance presented in the median plane and an interference signal presented at 30° (redrawn from [53]). (a) Spectrogram of the clean speech utterance. (b) Spectrogram of the mixture. (c) The ideal binary mask. (d) The estimated binary mask.

Fig. 14.2 shows a comparison between an ideal and an estimated T-F binary mask. Figs. 14.2(a) and (b) show the spectrograms of a clean speech utterance and the noisy mixture, respectively. The mixture is obtained using the spatial configuration in Fig. 14.1 and a factory noise as interference. The SNR is 0 dB. The algorithm described above is applied and the T-F binary mask obtained is shown in Fig. 14.2(d). Fig. 14.2(c) shows the corresponding ideal binary mask. As seen in Sec. 14.5, evaluations across a range of SNRs show that the estimated masks approximate very well the ideal binary mask under noisy but anechoic conditions. Similar results are obtained in [49] where the processing of ITD/IID statistics follows the description above but an auditory filterbank is used for frequency decomposition.

In reverberant conditions, the anechoic modeling of time delayed and attenuated mixtures is inadequate. Since the binaural cues of ITD and IID are smeared by reflections, the system performance of ITD/IID based binaural systems degrades considerably. A model of precedence effect is typically employed to improve the robustness against these smearing effects. The system proposed in [43] includes a delayed inhibition circuit which gives more weight to the onset of a sound in order to detect reliable spectral regions that are not contaminated by interfering noise or echoes. Speech recognition is then performed in the log spectral domain by employing missing data ASR. In order to account for the reverberant environment, a spectral energy normalization is employed before recognition. Similarly, the system proposed in [13] uses the interaural coherence to identify the T-F regions that are dominated by the direct sound. Soft masks are derived using probability distributions estimated from histograms of ITD/ILD estimates. The soft masks are then used as front-ends to a modified missing-data ASR. Under mildly reverberant conditions, the authors show that these techniques can improve ASR performance considerably.

14.4 Reverberant Binaural Segregation

We present here an alternative strategy to the binaural processors described previously which can deal more effectively with multiple interfering sources under reverberant conditions. The system proposed is a two-stage model that combines target cancellation with a nonlinear processing stage in order to estimate the ideal binary mask [50]. As seen in Fig. 14.3, an adaptive filter is applied in the first stage to the mixture signal, which contains both target and interference, in order to cancel the target signal. The adaptive filter is trained for simplification in the absence of noise. In the second stage, the system labels 1 only the T-F units that have been largely attenuated in the first stage since those units are likely to have originated from the target source; and labels 0 the other units.

The signal model in Fig. 14.3 assumes that a desired speech source $s(n)$ has been produced in a reverberant enclosure and recorded by two microphones

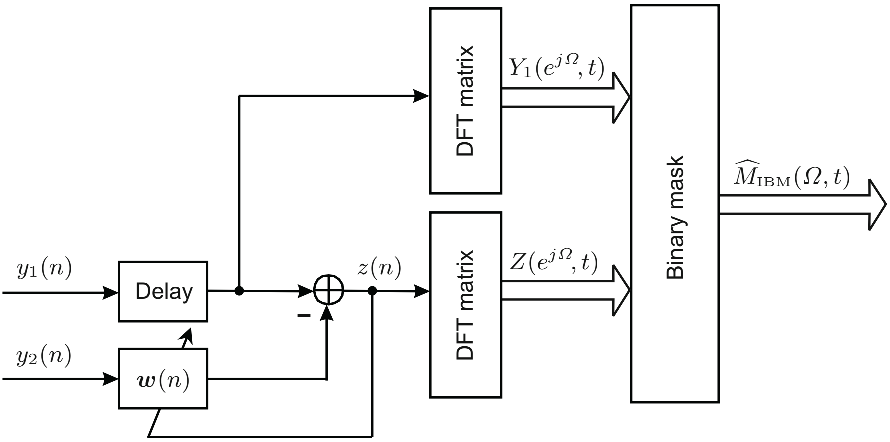

Fig. 14.3. Schematic diagram of the proposed model. The input signal is a mixture of reverberant target sound and acoustic interference. At the core of the system is an adaptive filter for target cancellation. The output of the system is an estimate of the ideal binary mask.

to produce the signal pair $[x_1(n), x_2(n)]$. . The transmission path from target location to microphones is a linear system and is modeled as:

$$x_1(n) = h_1(n) * s(n), \tag{14.5}$$
$$x_2(n) = h_2(n) * s(n), \tag{14.6}$$

where $h_i(n)$ corresponds to the room impulse response for the i'th microphone. The challenge of source separation arises when an unwanted interference pair $[n_1(n), n_2(n)]$ is also present at the input of the microphones resulting in a pair of mixtures $[y_1(n), y_2(n)]$:

$$y_1(n) = x_1(n) + n_1(n), \tag{14.7}$$
$$y_2(n) = x_2(n) + n_2(n). \tag{14.8}$$

The interference is a combination of multiple reverberant sources and additional background noise. Here, 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 target source location.

In the classical adaptive beamforming approach with two microphones [24], 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, the adaptive filter is used here for target cancellation. The noise reference is then used in a nonlinear scheme to estimate the ideal binary mask. This approach offers a potential solution to the multiple interference problem in reverberation.

In the experiments reported here, we assume a fixed target location and the filter $\boldsymbol{w}(n)$ 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 [25] with an impulse response of 375 ms length (6000 samples at 16 kHz sampling rate). 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 $z(n)$ and the noisy mixture at the primary microphone $y_1(n)$ 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 DFT 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:

$$OIR(\Omega, t) = \frac{\left|Z\left(e^{j\Omega}, t\right)\right|^2}{\left|Y_1\left(e^{j\Omega}, t\right)\right|^2}, \tag{14.9}$$

where $Y_1(e^{j\Omega}, t)$ and $Z(e^{j\Omega}, t)$ are the corresponding Fourier transforms of $y_1(n)$ and $z(n)$, respectively.

Consider a T-F unit in which the noise signal 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 $OIR(\Omega, t) \gg 0$. 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 $\widehat{M}_{\mathrm{IBM}}(\Omega, t)$ is 1 in those T-F units where $OIR(\Omega, t) > \theta(\Omega)$ and 0 in all the other units.

Fig. 14.4 shows a scatter plot of $R(\Omega, t)$ and $OIR(\Omega, t)$ obtained for individual T-F units corresponding to a frequency bin at 1 kHz. Similar results are seen across all frequencies. 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 (see Sec. 14.5 for simulation details); 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. 14.1. Fig. 14.5 demonstrates the performance of the proposed system for the following male target utterance: "Bright sunshine shimmers on the ocean" mixed with four interfering speakers at different locations. Observe that the estimated mask is able to estimate well the ideal binary mask especially in the high target energy T-F regions.

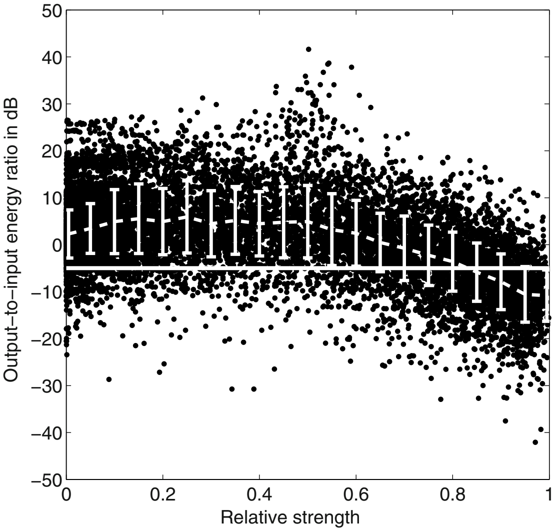

Fig. 14.4. Scatter plot of the output-to-input ratio with respect to the relative strength of the target to the mixture for a frequency bin centered at 1 kHz (from [50]). The mean and the standard deviation are shown as the dashed line and vertical bars, respectively. The horizontal line corresponds to the -6 dB decision threshold used in the binary mask estimation.

14.5 Evaluation

With the emergence of voice-based technologies, current ASR systems are required to deal with adverse conditions including noisy background and reverberation. Conventional ASR systems are constructed as a classification problem which involves the maximization of the posterior probability

$$p\big(W \,\big|\, \boldsymbol{Y}_{\text{sqr}}(t)\big),$$

where $\boldsymbol{Y}_{\text{sqr}}(t)$ is an observed short-term speech spectral power vector

$$\boldsymbol{Y}_{\text{sqr}}(t) = \Big[Y_{\text{sqr}}(\Omega = 0, t), \ldots, Y_{\text{sqr}}(\Omega = \pi, t)\Big]^{\text{T}} \tag{14.10}$$

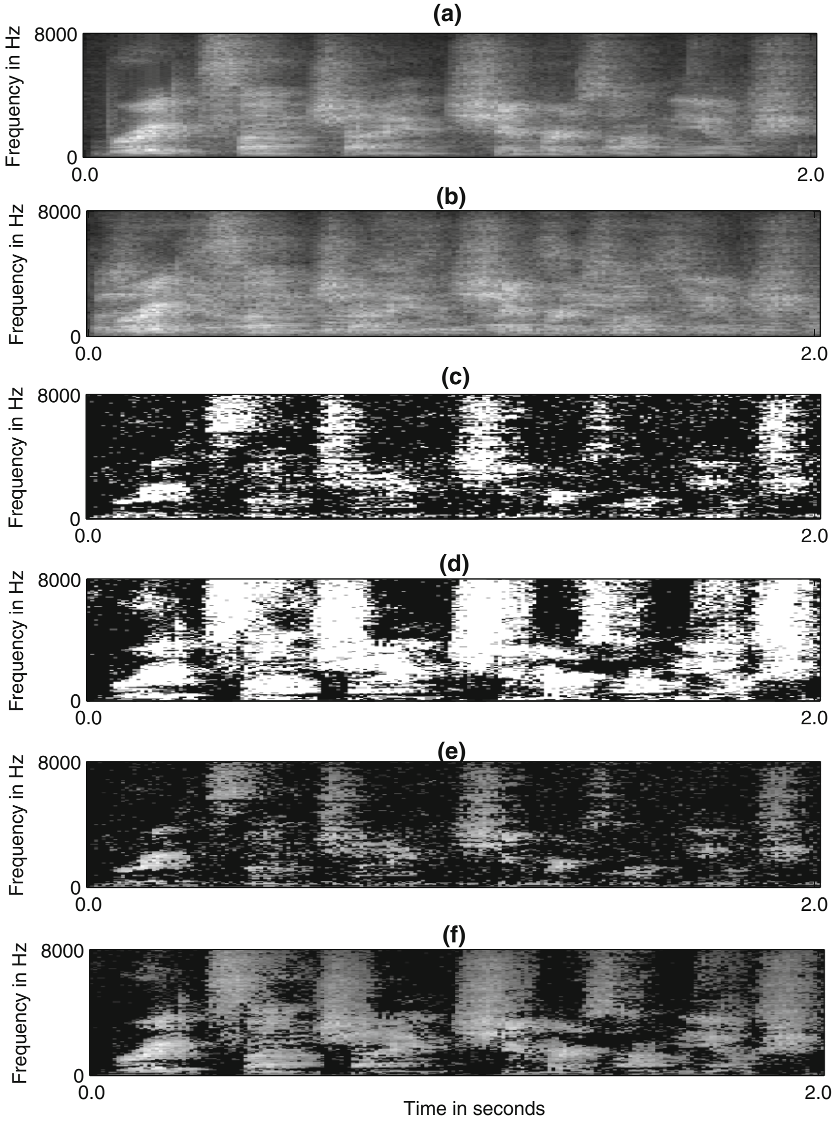

Fig. 14.5. A comparison between the estimated mask and the ideal binary mask for a five-source configuration (from [50]). (a) Spectrogram of the reverberant target speech. (b) Spectrogram of the mixture of target speech presented at 0° and four interfering speakers at locations −135°, −45°, 45° and 135°. The SNR is 5 dB. (c) The estimated T-F binary mask. (d) The ideal binary mask. (e) The mixture spectrogram overlaid by the estimated T-F binary mask. (f) The mixture spectrogram overlaid by the ideal binary mask. The recordings correspond to the left microphone.

with

$$Y_{\mathrm{sqr}}(\Omega, t) = \left| Y\left(e^{j\Omega}, t\right) \right|^2 \tag{14.11}$$

and W is a valid word sequence. The classification is highly sensitive to distortions in the spectral vector $\boldsymbol{Y}_{\mathrm{sqr}}(t)$. The standard approach to improving the ASR robustness is to enhance the target speech in the acoustic input. Given a T-F mask, the signal is resynthesized through the mask to reconstruct the target speech and the output is then fed to a conventional ASR system.

An alternative approach is the missing-data ASR proposed by Cooke et al. [17] which identifies the corrupted T-F spectral regions and treats them as unreliable or missing. In this approach, the spectral vector $\boldsymbol{Y}_{\mathrm{sqr}}(t)$ is partitioned into its reliable and unreliable components as $\boldsymbol{Y}_{\mathrm{sqr,r}}(t)$ and $\boldsymbol{Y}_{\mathrm{sqr,u}}(t)$, where $\boldsymbol{Y}_{\mathrm{sqr}}(t) = \boldsymbol{Y}_{\mathrm{sqr,r}}(t) \cup \boldsymbol{Y}_{\mathrm{sqr,u}}(t)$. The Bayesian decision is then sought using only the reliable components. In the marginalization method, the posterior probability is computed by integrating over the unreliable ones. However, further information about the mixing process can give lower and upper bounds for these unreliable components which can be used in the integral involved in marginalization. Under the assumption of additive and uncorrelated sound sources, the true value of the speech energy in the unreliable parts can be constrained between 0 and the observed spectral energy $\boldsymbol{Y}_{\mathrm{sqr,u}}(t)$. The T-F units indicated as 1 in the binary mask are the reliable units while those indicated as 0 are the unreliable ones. It has been shown that this approach outperforms conventional ASRs with input resynthesized from T-F binary masks. Moreover, ideal binary masks produce impressive recognition scores when applied to the missing-data ASR for a variety of noise intrusions including multiple interfering sources [17, 49].

The binaural system presented in Sec. 14.3 has been evaluated under noisy but anechoic conditions using the missing-data ASR. As in [17], the task domain is speaker independent recognition of connected digits. Thirteen (the number 1-9, a silence, very short pause between words, zero and oh) word-level models are trained using an hidden Markov model (HMM) toolkit, HTK [63]. 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 grammar for this task allows for one or more repetitions of digits and all digits are equally probable. Both training and testing are performed using the male speaker dataset in the TIDigits database [34]. 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 are sampled at 20 kHz.

Fig. 14.6 shows recognition results for target source in the median plane and one noise source on the right side at 30° for a range of SNRs from -5 to 10 dB. The noise source is the factory noise from the NOISEX corpus [57]. The factory noise is chosen as it has energy in the formant regions, therefore posing challenging problems for recognition. The binaural mixtures are obtained by convolving the original signals with the HRIRs of the corresponding sound source locations. The binaural processing described in Sec. 14.3 is applied to derive the corresponding T-F binary mask which is then fed to the missing-data ASR. Feature vectors for the missing-data ASR are derived from the 512 DFT coefficients extracted in each time frame. Recognition is performed using log-spectral energy bandlimited to 4 kHz. Hence only 98 spectral coefficients along with delta coefficients in a two-frame delta-window are extracted in each frame. As seen in Fig. 14.6, the estimated masks approximate very well the ideal binary masks resulting in large recognition improvements over the baseline. Similar results have been obtained in multispeaker conditions in [49].

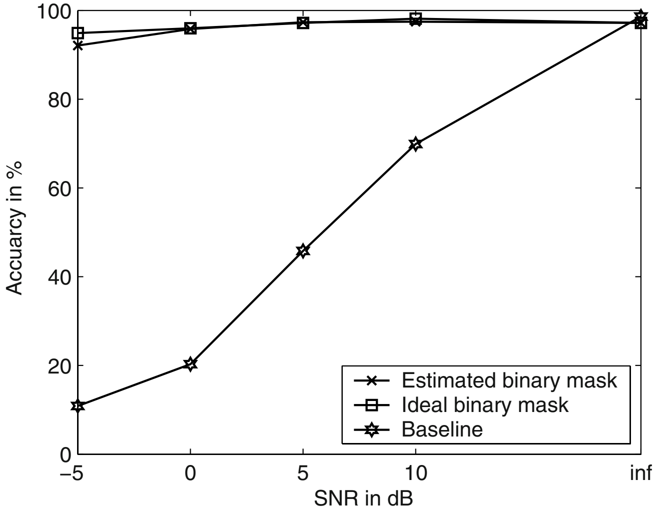

Fig. 14.6. Comparison between estimated and ideal binary masks as front-ends to a missing-data ASR under anechoic conditions (redrawn from [53]). The target source in the median plane is presented with a noise source on the right side at 30° for a range of SNRs from −5 to 10 dB. For comparison, the baseline performance is shown.

To illustrate the binaural system described in Sec. 14.4, we present here systematic recognition results under multi-source reverberant conditions. The reverberation is generated using the room acoustic model described in [43]. The reflection paths of a particular sound source are obtained using the image reverberation model for a small rectangular room (6 m × 4 m × 3 m) [1]. The resulting impulse response is convolved with the same HRIRs as before in order to produce the binaural input 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 (2.5 m × 2.5 m × 2 m) 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:

- an interference of rock music at 45° (scene 1);
- two concurrent speakers (one female and one male utterance) at azimuth angles of −45° and 45° (scene 2); and
- 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 conditions scene 2 and scene 3 making them more comparable with condition scene 1. The signals are upsampled to 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.

While the missing-data approach has shown promising results with additive noise in anechoic conditions, an extension to reverberant conditions has turned out to be problematic (see for example [43]). We therefore adapt here the spectrogram reconstruction method proposed in [46] to reverberant conditions which shows improved performance over the missing-data recognizer. This approach has been suggested in the context of additive noise. In this approach, a noisy spectral vector $\boldsymbol{Y}_{\mathrm{sqr}}(t)$ at a particular frame is partitioned in its reliable $\boldsymbol{Y}_{\mathrm{sqr,r}}(t)$ and its unreliable $\boldsymbol{Y}_{\mathrm{sqr,u}}(t)$ components. The task is to reconstruct the underlying true spectral vector $\boldsymbol{X}_{\mathrm{sqr}}(t)$. Assuming that the reliable features $\boldsymbol{Y}_{\mathrm{sqr,r}}(t)$ are approximating well the true ones $\boldsymbol{X}_{\mathrm{sqr,r}}(t)$, a Bayesian decision is then employed to estimate the remaining $\boldsymbol{X}_{\mathrm{sqr,u}}(t)$ 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. Although the reliable data in our system contains some reverberation, we train the prior speech model only on clean data. This actually avoids the trouble of obtaining a prior for each deployment condition, and is desirable for robust speech recognition. The signals reconstructed in this way are then used as input to a conventional ASR which employs cepstral features.

The speech prior is modeled empirically as a mixture of Gaussians and trained on the clean database used for training the conventional ASR (see below):

$$p\big(\boldsymbol{X}_{\mathrm{sqr}}(t)\big) = \sum_{k=1}^{M} p(k)p\Big(\boldsymbol{X}_{\mathrm{sqr}}(t)\big|\, k\Big), \tag{14.12}$$

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

Previous studies [17, 46] have shown that a good estimate of $\boldsymbol{X}_{\mathrm{sqr,u}}(t)$ is its mean conditioned on $\boldsymbol{X}_{\mathrm{sqr,r}}(t)$:

$$\begin{aligned} &\mathrm{E}\Big\{\boldsymbol{X}_{\mathrm{sqr,\,u}}(t)\big|\,\boldsymbol{X}_{\mathrm{sqr,r}}(t),\, 0 \le \boldsymbol{X}_{\mathrm{sqr,u}}(t) \le \boldsymbol{Y}_{\mathrm{sqr,u}}(t)\Big\} \\ &= \sum_{k=1}^{M} p\big(k\,\big|\,\boldsymbol{X}_{\mathrm{sqr,r}}(t),\, 0 \le \boldsymbol{X}_{\mathrm{sqr,u}}(t) \le \boldsymbol{Y}_{\mathrm{sqr,u}}(t)\big) \\ &\quad \underbrace{\cdot \int_{0}^{\boldsymbol{Y}_{\mathrm{sqr,u}}(t)} X\, p\big(X\,\big|\,k, 0 \le X \le \boldsymbol{Y}_{\mathrm{sqr,u}}(t)\big)\, dX}_{\widetilde{\boldsymbol{X}}_{\mathrm{sqr,u}}(t)} \end{aligned} \tag{14.13}$$

where $p(k\,|\,\boldsymbol{X}_{\mathrm{sqr,r}}(t), ...)$ is the *a posteriori* probability of the k'th Gaussian given the reliable data and the integral denotes the expectation $\widetilde{\boldsymbol{X}}_{\mathrm{sqr,u}}(t)$ corresponding to the k'th mixture. Note that under the additive noise condition, the unreliable parts may be constrained as $0 \le \boldsymbol{X}_{\mathrm{sqr,u}}(t) \le \boldsymbol{Y}_{\mathrm{sqr,u}}(t)$ [17]. Here it is 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, 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:

$$\widetilde{\boldsymbol{X}}_{\mathrm{sqr,u}}(t) = \begin{cases} \mu_{\mathrm{u},k} & , \;\; 0 \le \mu_{\mathrm{u},k} \le \boldsymbol{Y}_{\mathrm{sqr,u}}(t), \\ \boldsymbol{Y}_{\mathrm{sqr,u}}(t) & , \;\; \mu_{\mathrm{u},k} > \boldsymbol{Y}_{\mathrm{sqr,u}}(t), \\ 0 & , \;\; \mu_{\mathrm{u},k} < 0. \end{cases} \tag{14.14}$$

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

The same recognition task is used as in the previous evaluation. Training is performed using the 4235 clean signals from the male speaker dataset in the

TIDigits database downsampled to 16 kHz to be consistent with the system described in Sec. 14.4. The HMMs are trained with clean utterances from the training data using feature vectors consisting of the 13 mel-frequency cepstral coefficients (MFCC) including the zeroth order cepstral coefficient, $C_0(n)$, as the energy term 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 [45]. Cepstral mean normalization (CMN) is applied to the cepstral features in order to improve the robustness of the system under reverberant conditions [52]. 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 result using clean test utterances is 99 % accuracy. Using the reverberated test utterances, performance degrades to 94 % accuracy.

Testing is performed on a subset of the testing set containing 229 utterances from 3 speakers which is similar to the test used in [43]. 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 conditions of Scene 1 to Scene 3. Speech recognition results for the three conditions are reported separately in Figs. 14.7, 14.8 and 14.9 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 (delay-and-sum),
- 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.

Note that the ASR performance depends on the interference type and we obtain the best accuracy score in the two speaker interference. The baseline results correspond to the unprocessed left ear signal. Observe that our system achieves large improvements over the baseline performance across all conditions.

The adaptive beamformer used in evaluations follows the two-stage adaptive filtering strategy described in [56] that improves the classic Griffiths-Jim model [24] 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 in order to extract the target signal. Here, training for the second filter is done independently for each noise condition 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). As seen in Fig. 14.7, the adaptive beamformer outperforms all the other algorithms in the case of

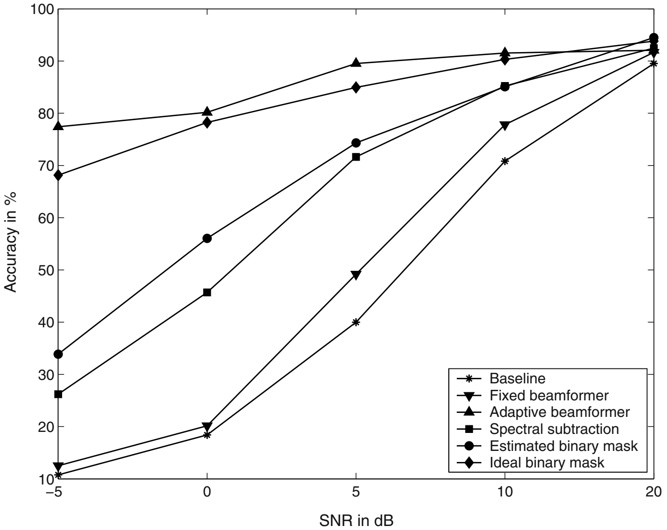

Fig. 14.7. Recognition performance for Scene 1 at different SNR values for the reverberant mixture ($*$), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), an ASR front-end using the estimated binary mask (●), and an ASR front-end using the ideal binary mask (♦) (from [50]).

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 scene 3. As proposed in [2], we can combine the target cancellation stage with spectral subtraction to attenuate the interference. As illustrated by the recognition results in Figs. 14.8 and 14.9, 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 ASR front-end employs a better estimation of the spectrum in these unreliable T-F units and therefore results in large improvements over the spectral subtraction method. Although the results using our ASR front-end show substantial performance gains, further improvement can be achieved as can be seen in the results reported with the ideal binary mask.

14.6 Concluding Remarks

In anechoic conditions, there exists a systematic relationship between computed ITD and IID values and local SNR within individual T-F units. This relationship leads to characteristic clustering in the joint ITD-IID feature

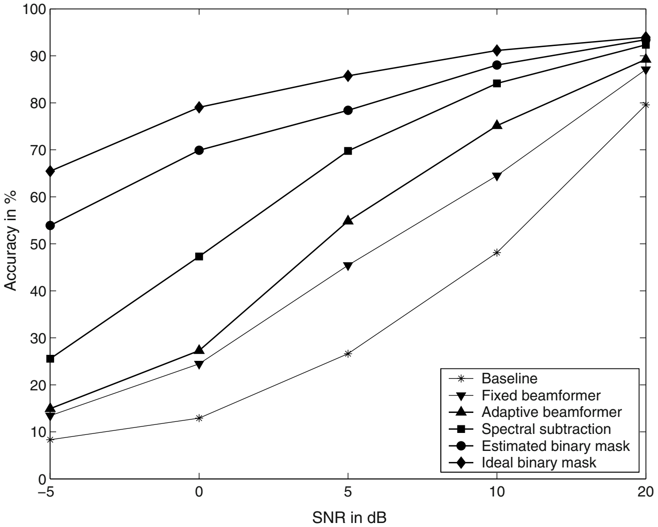

Fig. 14.8. Recognition performance for Scene 2 at different SNR values for the reverberant mixture (∗), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), an ASR front-end using the estimated binary mask (•), and an ASR front-end using the ideal binary mask (⧫) (from [50]).

space, which enables the effective use of supervised classification to estimate the ideal binary mask. Estimated binary masks thus obtained from mixtures of target speech and acoustic interference have been shown to match the ideal ones very well.

In natural settings, reverberation alters many of the acoustical properties of a sound source reaching our ears, including smearing 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 binaural 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 highly affected by these distortions. In this chapter we have described strategies to alleviate this problem as well as a system that integrates target cancellation through adaptive filtering and T-F binary masking which is able to perform well under multi-source reverberant conditions.

Most work in binaural CASA assumes that sound sources remain fixed throughout testing. The system proposed in Sec. 14.4 alleviates somehow the problem; it is insensitive to interference location changes but assumes a fixed target location. None of these are realistic situations since head movement as well as source movement can occur. One way to approach the problem is to add a source tracking component. For example, the system proposed in [48] is able to track the azimuths of multiple acoustic sources using ITD/IID estimates.

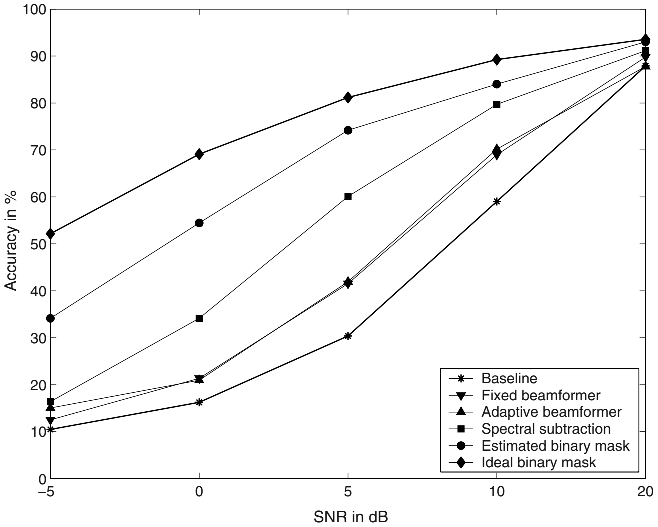

Fig. 14.9. Recognition performance for Scene 3 at different SNR values for the reverberant mixture (∗), a fixed beamformer (▼), an adaptive beamformer (▲), a system that combines target cancellation and spectral subtraction (■), an ASR front-end using the estimated binary mask (•), and an ASR front-end using the ideal binary mask (⧫) (from [50]).

Such a system could be coupled with a binaural processor to deal with moving sources. Nix and Hohmann [42] have recently proposed to simultaneously track sound source locations and spectral envelopes using a non-Gaussian multidimensional statistical filtering approach. This strategy could be used to integrate in a robust way different acoustic cues even when they are corrupted by noise, reverberation and motion. Other two-microphone algorithms combine location cues and pitch information or other signal processing techniques to improve system robustness [4, 36, 51].

The computational goal of many CASA algorithms is the ideal binary T-F mask which selects target-dominant spectrotemporal regions. Signals reconstructed from such masks have been shown to be substantially more intelligible for human listeners than the original mixtures [14, 49]. However, conventional ASR systems are extremely sensitive to the distortions produced during resynthesis. Here, we have utilized two strategies that minimize these effects on recognition:

- the missing-data ASR proposed by Cooke et al. [17] that utilizes only the reliable target dominant features in the acoustic mixture
- and a target reconstruction method for the unreliable features proposed by Raj et al. [46].

As seen in our evaluations, the proposed binaural CASA systems coupled with these two strategies can produce substantial ASR improvements over baseline under both anechoic and reverberant multi-source conditions. Recently, a new approach to robust speech recognition has been proposed to additionally take into account the varied accuracy of features derived from front-end preprocessing [18]. Srinivasan and Wang [55] convert binary uncertainty in the T-F mask into real-valued uncertainty associated with cepstral features, which can then be used by an uncertainty decoder during recognition. Such uncertainty-based strategies can be utilized to further improve the performance of current binaural CASA systems when applied to robust speech recognition [54].

14.7 Acknowledgments

This research described in this chapter was supported in part by an AFOSR grant (FA9550-04-01-0117) and an NSF grant (IIS-0534707).

References

1. J. B. Allen, D. A. Berkley: Image method for efficiently simulating small-room acoustics, *JASA* , **65**, 943–950, 1979.
2. A. Álvarez, P. Gómez, V. Nieto, R. Martínez, V. Rodellar: Speech enhancement and source separation supported by negative beamforming filtering, *Proc. ICSP '02,* 342–345, 2002.
3. S. Araki, S. Makino, H. Sawada, R. Mukai: Underdetermined blind separation of convolutive mixtures of speech with directivity pattern based mask and ICA, *Proc. Fifth International Conference on Independent Component Analysis 04*, 898-905, 2004.
4. A. K. Barros, T. Rutkowski, F. Itakura, N. Ohnishi: Estimation of speech embedded in a reverberant and noisy environment by independent component analysis and wavelets, *IEEE Trans. Neural Netw.,* **13**, 888–893, 2002.
5. R. Balan, A. Jourjine, J. Rosca: AR processes and sources can be reconstructed from degenerate mixtures, *Proc. 1st International Workshop on Independent Component Analysis and Signal Separation,* 467–472, 1999.
6. J. Blauert: *Spatial Hearing – The Psychophysics of Human Sound Localization,* Cambridge, MA, USA: MIT press, 1997.
7. M. Bodden: Modeling human sound-source localization and the cocktail-party-effect, *Acta Acoustica,* **1**, 43–55, 1993.
8. M. Brandstein, D. Ward (eds.): *Microphone Arrays: Signal Processing Techniques and Application,* Berlin, Germany: Springer, 2001.
9. A. S. Bregman: *Auditory Scene Analysis,* Cambridge, MA, USA: MIT press, 1990.
10. A. W. Bronkhorst: The cocktail party phenomenon: a review of research on speech intelligibility in multiple-talker conditions, *Acustica,* **86**, 117–128, 2000.
11. G. J. Brown, M. P. Cooke: Computational auditory scene analysis, *Comput. Speech Lang.,* **8**, 297–336, 1994.
12. G. J. Brown, D. L. Wang: Separation of speech by computational auditory scene analysis, in J. Benesty, S. Makino, J. Chen (eds.), *Speech Enhancement,* New York, NY, USA: Springer, 2005, 371–402.

13. G. J. Brown, S. Harding, J. P. Barker: Speech separation based on the statistics of binaural auditory features, *Proc. ICASSP '06,* **5**, 2006.
14. D. S. Brungart, P. S. Chang, B. D. Simpson, D. L. Wang: Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation, *JASA,* **120**, 4007–4018, 2006.
15. E. C. Cherry: Some experiments on the recognition of speech, with one and with two ears, *JASA,* **25**, 975–979, 1953.
16. M. P. Cooke: *Modeling Auditory Processing and Organization,* Cambridge, U.K.: Cambridge University Press, 1993.
17. M. P. Cooke, P. Green, L. Josifovski,A. Vizinho: Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Commun.,* **34**, 267–285, 2001.
18. L. Deng, J. Droppo, A. Acero: Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion, *IEEE Trans. Speech, and Audio Process.,* **13**, 412-421, 2005.
19. P. Divenyi (ed.): *Speech Separation by Humans and Machines,* Norwell, MA, USA: Kluwer Academic, 2005.
20. Y. Ephraim, H. L. Trees: A signal subspace approach for speech enhancement, *IEEE Trans. Speech Audio Process.,* **3**, 251–266, 1995.
21. Y. Ephraim, D. Malah: Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, *IEEE Trans. Acoust. Speech Signal Process.,* **ASSP-32**(6), 1109-1121, 1984.
22. W. G. Gardner, K. D. Martin: HRTF measurements of a KEMAR dummy-head microphone, *MIT Media Lab Perceptual Computing Technical Report #280,* 1994.
23. J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren: Darpa timit acoustic-phonetic continuous speech corpus, *Technical Report NISTIR 4930,* National Institute of Standards and Technology, Gaithersburg, MD, USA, 1993.
24. L. J. Griffiths, C. W Jim: An alternative approach to linearly constrained adaptive beamforming, *IEEE Trans. Antennas and Propagation,* **30**, 27–34, 1982.
25. S. Haykin: *Adaptive Filter Theory*, 4th ed., Upper Saddle River, NJ, USA: Prentice Hall, 2002.
26. H. Helmholtz: *On the Sensation of Tone*, (A. J. Ellis, Trans.), 2nd English ed., New York, NY, USA: Dover Publishers, 1863.
27. G. Hu, D. L. Wang: Monaural speech segregation based on pitch tracking and amplitude modulation, *IEEE Trans. Neural Netw.,* **15**, 1135–1150, 2004.
28. G. Hu, D. L. Wang: An Auditory Scene Anaylsis Approach to Monaural Speech Segregation, in E. Hänsler, G. Schmidt (eds.), *Topis in Acoustic Echo and Noise Control*, 485–515, Berlin, Germany: Springer, 2006.
29. G. Hu, D. L. Wang: Auditory segmentation based on onset and offset analysis, *IEEE Trans. Audio, Speech and Language Process.,* **15**, 396–405, 2007.
30. X. Huang, A. Acero, H. -W. Hon: *Spoken Language Processing: A guide to theory, algorithms, and system development,* Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.
31. A. Hyväarinen, J. Karhunen, E. Oja: *Independent component analysis,* New York, NY, USA: Wiley, 2001.
32. L. A. Jeffress: A place theory of sound localization, *Journal of Comparative and Physiological Psychology,* **41**, 35–39, 1948.

33. R. P. Lippman: Speech recognition by machines and humans, *Speech Commun.,* **22**, 1–16, 1997.
34. R. G. Leonard: A database for speaker-independent digit recognition, *Proc. ICASSP '84,*, 111–114, 1984.
35. C. Liu, B. C. Wheeler, W. D. O'Brien, Jr., C. R. Lansing, R. C. Bilger, D. L. Jones, A.S. Feng: A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers, *JASA,* **110**, 3218–3230, 2001.
36. H. Y. Luo, P. N. Denbigh: A speech separation system that is robust to reverberation, *Proc. International Symposium on Speech, Image Process. and Neural Netw.,* 339–342, 1994.
37. R. F. Lyon: A computational model of binaural localization and separation, *Proc. ICASSP '83,* 1148–1151, 1983.
38. N. Ma, M. Bouchard, R. Goubran: Perceptual Kalman filtering for speech enhancement in colored noise, *Proc. ICASSP '04,* **1**, 717–720, 2004.
39. E. A. MacPherson: A computer model of binaural localization for stereo imaging measurement, *J. Audio Engineering Soc.,* **39**, 604–622, 1991.
40. R. Martin: Noise power spectral density estimation based on optimal smoothing and minimum statistics, *IEEE Trans. Speech Audio Process.,* **9**, 504–512, 2001.
41. B. C. J. Moore: *An introduction to the Psychology of Hearing,* 5th ed., San Diego, CA, USA: Academic, 2003.
42. J. Nix, V. Hohmann: Combined estimation of spectral envelopes and sound source direction ofconcurrent voices by multidimensional statistical filtering, *IEEE Trans. Audio, Speech and Language Process.,* **15**, 995–1008, 2007.
43. K. J. Palomäki, G. J. Brown, D. L. Wang: A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation, *Speech Commun.,* **43**, 361–378, 2004.
44. M. S. Pedersen, D. L. Wang, J. Larsen, U. Kjems: Two-microphone separation of speech mixtures,*IEEE Trans. Neural Netw.,* in press, 2008.
45. L. R. Rabiner, B. H. Juang: *Fundamentals of Speech Recognition,* 2nd ed., Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
46. B. Raj, M. L. Seltzer, R. M. Stern: Reconstruction of missing features for robust speech recognition, *Speech Commun.,* **43**, 275–296, 2004.
47. N. Roman, D. L. Wang, G. J. Brown: Speech segregation based on sound localization, *Proc. IJCNN '01,* 2861–2866, 2001.
48. N. Roman, D. L. Wang: Binaural tracking of multiple moving sources, *Proc. ICASSP '03,* **5**, 149–152, 2003.
49. N. Roman, D. L. Wang, G. J. Brown: Speech segregation based on sound localization, *JASA,* **114**, 2236–2252, 2003.
50. N. Roman, S. Srinivasan, D. L. Wang: Binaural segregation in multisource reverberant environments, *JASA,* **120**, 4040–4051, 2006.
51. A. Shamsoddini, P. N. Denbigh: A sound segregation algorithm for reverberant conditions, *Speech Commun.,* **33**, 179–196, 2001.
52. M. L. Shire: 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, 2000.
53. S. Srinivasan, N. Roman, D. L. Wang: Binary and ratio time-frequency masks for robust speech recognition, *Speech Commun.,* **48**, 1486–1501, 2006.
54. S. Srinivasan, N. Roman, D. L. Wang: Exploiting uncertainties for binaural speech recognition, *Proc. ICASSP '07,* **4**, 789–792, 2007.

55. S. Srinivasan, D. L. Wang: Transforming binary uncertainties for robust speech recognition, *IEEE Trans. Audio, Speech, and Language Process.,* **15**, 2130–2140, 2007.
56. D. Van Compernolle: Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings, *Proc. ICASSP '90,* 833–836, 1990.
57. A. P. Varga, H. J. M. Steeneken,M. Tomlinson, D. Jones: The NOISEX-92 study on the effect of additive noise on automatic speech recogonition, Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK, 1992.
58. D. L. Wang: On ideal binary mask as the computational goal of auditory scene analysis, in P. Divenyi (ed.), *Speech Separation by Humans and Machines,* Norwell, MA, USA: Kluwer Academic, 2005, 181–197.
59. D. L. Wang, G. J. Brown: Separation of speech from interfering sounds based on oscillatory correlation, *IEEE Trans. Neural Netw.,* **10**, 684–697, 1999.
60. D. L. Wang, G. J. Brown (eds.): *Computational auditory scene analysis: Principles, algorithms and applications,* IEEE Press/Wiley-Interscience, 2006.
61. T. Whittkop, V. Hohmann: Strategy-selective noise reduction for binaural digital hearing aids,*Speech Commun.*, **39**, 111–138, 2003.
62. O. Yilmaz, S. Rickard: Blind separation of speech mixtures via time-frequency masking, *IEEE Trans. Signal Process.,* **52**(7), 1830–1847, 2004.
63. S. Young, D. Kershaw, J. Odell, V. Valtchev, P. Woodland: The HTK Book (for HTK Version 3.0), Microsoft Corporation, 2000.
64. M. Zibulevsky, B. A. Pearlmutter, P. Bofill, P. Kisilev: Blind source separation by sparse decomposition, in S. J. Roberts, R. M. Everson (eds.), *Independent Component Analysis: Principles and Practice,* Cambridge University Press, 2001.