16

Computational Auditory Scene Analysis and Automatic Speech Recognition

Arun Narayanan, DeLiang Wang
The Ohio State University, USA

16.1 Introduction

The human auditory system is, in a way, an engineering marvel. It is able to do wonderful things that powerful modern machines find extremely difficult. For instance, our auditory system is able to follow the lyrics of a song when the input is a mixture of speech and musical accompaniments. Another example is a party situation. Usually there are multiple groups of people talking, with laughter, ambient music and other sound sources running in the background. The input our auditory system receives through the ears is a mixture of all these. In spite of such a complex input, we are able to selectively listen to an individual speaker, attend to the music in the background, and so on. In fact this ability of 'segregation' is so instinctive that we take it for granted without wondering about the complexity of the problem our auditory system solves.

Colin Cherry, in the 1950s, coined the term 'cocktail party problem' while trying to describe how our auditory system functions in such an environment [12]. He did a series of experiments to study the factors that help humans perform this complex task [11]. A number of theories have been proposed since then to explain the observations made in those experiments [11,12,70]. Helmhotz had, in the mid-nineteenth century, reflected upon the complexity of this signal by using the example of a ball room setting [22]. He remarked that even though the signal is "complicated beyond conception," our ears are able to "distinguish all the separate constituent parts of this confused whole."

So how does our auditory system solve the so-called cocktail party problem? Bregman tried to give a systematic account in his seminal 1990 book *Auditory Scene Analysis* [8]. He calls

Techniques for Noise Robustness in Automatic Speech Recognition, First Edition.
Edited by Tuomas Virtanen, Rita Singh, and Bhiksha Raj.

the process "scene analysis" by drawing parallels with vision. It has been argued that the goal of perception is to form a mental description of the world around us. Our brain analyzes the scene and forms mental representations by combining the evidence that it gathers through the senses. The role of audition is no different. Its goal is to form a mental description of the *acoustic* world around us by integrating sound components that belong together (e.g., those of the target speaker in a party) and segregating those that do not. Bregman suggests that the auditory system accomplishes this task in two stages. First, the acoustic input is broken down into local time-frequency elements, each belonging to a single source. This stage is called segmentation as it forms locally grouped time-frequency regions or *segments* [79]. The second stage then groups the segments that belong to the same source to form an *auditory stream*. A stream corresponds to a single source.

Inspired by Bregman's account of auditory organization, many computational systems have been proposed to segregate sound mixtures automatically. Such algorithms have important practical applications in hearing aids, automatic speech recognition, automatic music transcription, etc. The field is collectively termed *Computational Auditory Scene Analysis* (CASA).

This chapter is about CASA and automatic speech recognition in noise. In Section 16.2, we discuss some of the grouping principles of auditory scene analysis (ASA), focusing primarily on the cues that are most important for the auditory organization of speech. We then move on to computational aspects. How to combine CASA and ASR effectively is, in itself, a research issue. We address this by discussing CASA in depth, and introducing an important goal of CASA - *Ideal Binary Mask* (IBM) - in Section 16.3. As we will see, the IBM has applications to both speech segregation and automatic speech recognition. We will also discuss a typical architecture of CASA systems in Section 16.3. This will be followed by a discussion of strategies used for IBM estimation in Section 16.4. In the subsequent section, we address the topic of robust automatic speech recognition, where we will discuss some of the methods to integrate CASA and ASR. We note that this topic will also be addressed in other chapters (see Chapters 14 and 15 for detailed descriptions on missing-data ASR techniques). Finally, Section 16.6 offers a few concluding remarks.

16.2 Auditory Scene Analysis

CASA-based systems use ASA principles as a foundation to build computational models. As mentioned in the introductory section, Bregman described ASA to be a two stage process which results in integration of acoustic components that belong together and segregation of those that do not. In the first stage, an acoustic signal is broken down into time-frequency (T-F) segments. The second stage groups segments formed in the first stage into streams. Grouping of segments can occur across frequency or across time. They are called *simultaneous grouping* and *sequential grouping*, respectively.

A number of factors influence the grouping stage which results in the formation of coherent streams from local segments. Two distinctive schemes have been described by Bregman: primitive grouping and schema-based grouping.

Primitive grouping is an innate bottom-up process that groups segments based on acoustic attributes of sound sources. Major primitive grouping principles include proximity, periodicity, continuity, common onset/offset, amplitude and frequency modulation, and spatial location [8, 79]. Proximity refers to closeness in time or frequency of sound components. The components of a periodic signal are harmonically related (they are multiples of the fundamental frequency or $F0$), and thus segments that are harmonically related are grouped together. Periodicity is a major grouping cue that has also been widely utilized by CASA systems. Continuity refers to the continuity of pitch (perceived fundamental frequency), spectral and temporal continuity, etc. Continuity or smooth transitions can be used to group segments across time. Segments that have synchronous onset or offset times are usually associated with the same source and hence, grouped together. Among the two, onset synchrony is a stronger grouping cue. Similarly, segments that share temporal modulation characteristics (amplitude or frequency) tend to be grouped together. If segments originate from the same spatial location, there is a high probability that they belong to the same source and hence should be grouped.

Unlike primitive grouping, schema-based grouping is a top-down process where grouping occurs based on the learned patterns of sound sources. Schema-based organization plays an important role in grouping segments of speech and music, as some of their properties are learned over time by the auditory system. An example is the identification of a vowel based on observed formants. Note that both schema-based and primitive grouping play important roles in organizing real-world signals like speech and music.

The grouping principles introduced thus far were originally found though laboratory experiments using simple stimuli such as tones. Later experiments using more complex speech stimuli have established their role in speech perception [2,8]. Figure 16.1 shows some of the primitive grouping cues present for speech organization. Cues like continuity, common onset/offset, harmonicity are marked in the figure.

16.3 Computational Auditory Scene Analysis

Wang and Brown define CASA as ([79], p. 11):

> . . . the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene.

This definition takes into account the biological relevance of this field by limiting the number of microphones to two (like in humans) and the functional goal of CASA. The mechanisms used by CASA systems are perceptually motivated. For example, most systems make use of harmonicity as a grouping cue [79]. But this does not mean that the systems are exclusively dependent on ASA to achieve their goals. As we will see, modern systems make use of perceptual cues in combination with methods not necessarily motivated from the biological perspective.

16.3.1 Ideal Binary Mask

The goal of ASA is to form perceptual streams corresponding to the sound sources from the acoustic signal that reaches our ears. Taking this into consideration, Wang and colleagues suggested the *Ideal Binary Mask* as a main goal of CASA [24,27,76]. The concept was largely motivated by the masking phenomenon in auditory perception, whereby a stronger sound masks a weaker sound and renders it inaudible within a critical band [49]. Along the same lines, the IBM defines what regions in the time-frequency representation of a mixture are target dominant and what regions are not. Assuming a spectrogram-like representation of an

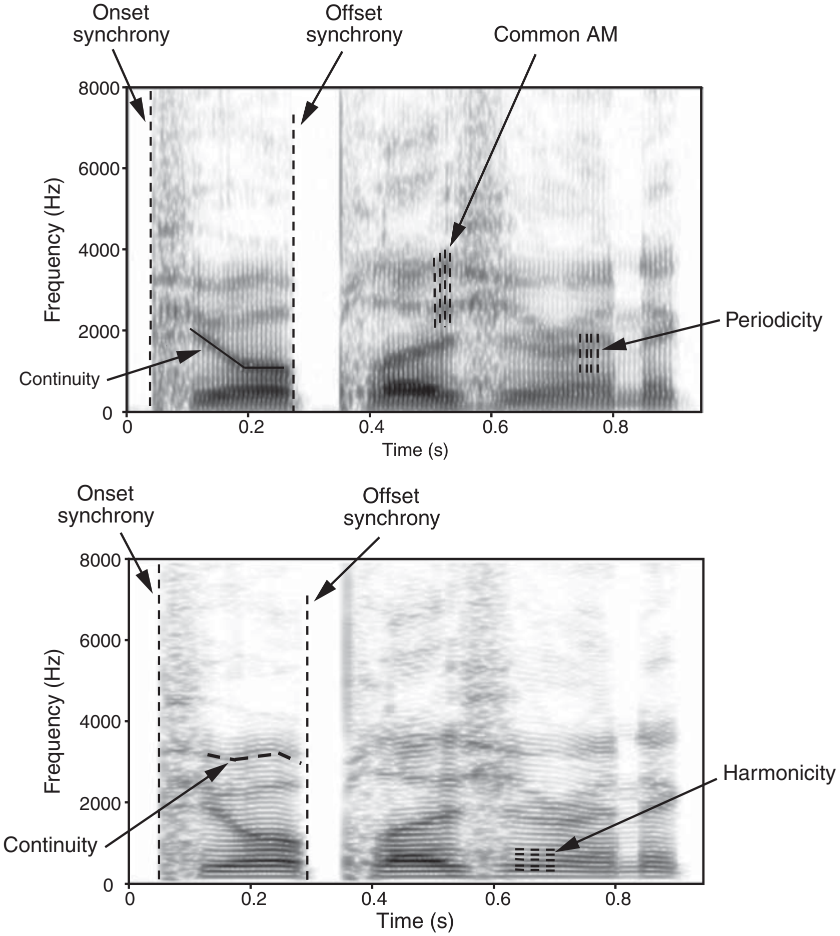

Figure 16.1 Primitive grouping cues for speech organization (reproduced from Wang and Brown [79]). The top panel shows a broadband spectrogram of the utterance "pure pleasure". Temporal continuity, onset and offset synchrony, common amplitude modulation and harmonicity cues are present. The bottom panel shows a narrow-band spectrogram of the same utterance.

acoustic input, the IBM takes the form of a binary matrix with 1 representing target dominant T-F units and 0 representing interference dominant units.

Mathematically, the IBM is defined as:

$$IBM(t,f) = \begin{cases} 1 & \text{if } SNR(t,f) \geq LC \\ 0 & \text{otherwise.} \end{cases} \tag{16.1}$$

Here, $SNR(t,f)$ represents the signal-to-noise ratio (SNR) within the T-F unit of time index t and frequency index (or channel) f. LC stands for a local criterion, which acts as an SNR threshold that determines how strong the target should be over the noise for the unit to be marked target dominant. The LC is usually set to 0 dB which translates to a simple rule of whether the target energy is stronger than the noise energy. Note that, to obtain the IBM, we need access to the premixed target and interference signals (hence the term "ideal"). According to them, a CASA system should aim at *estimating* the IBM from the mixture signal. It should be pointed out that the IBM can be thought of as an "oracle" binary mask. Oracle masks, binary or otherwise, have been widely used in the missing-data ASR literature to indicate the ceiling recognition performance of noisy speech.

The reasons why the IBM is an appropriate goal of CASA include the following:

(i) Li and Wang studied the optimality of the IBM measured in terms of the improvement in the SNR of a noisy signal (SNR gain) processed using binary masks [43]. They show that, under certain conditions, the IBM with the *LC* of 0 dB is optimal among all binary masks. Further, they compare the IBM with the ideal ratio (soft) mask, which is a T-F mask with real values representing the percentages of target speech energy contained in T-F units, similar to a Wiener filter. The comparisons show that, although the ideal ratio mask achieves higher SNR gains than the IBM as expected, in most mixtures of interest the difference in SNR gain is very small.

(ii) IBM-segregated noisy speech has been shown to greatly improve intelligibility for both normal hearing and hearing impaired listeners [1,10,42,81]. Even when errors are introduced to the IBM, it can still improve the intelligibility of noisy speech as long as the errors are within a reasonable range [42,62]. Moreover, it has been found that the *LC* of –6 dB seems to be more effective than the *LC* of 0 dB to improve speech intelligibility [81] even though the latter threshold leads to a higher SNR of IBM processed signals.

(iii) Speech energy is sparsely distributed in a high-resolution T-F representation, and there is little overlap between the components of different speakers in a speech mixture [63,86]. Under such circumstances, the IBM can almost segregate a mixture into its constituent streams. Note that sparsity does not hold for broadband interferences such as speech babble or when room reverberation is present.

(iv) Related binary masks have been shown to be effective for robust ASR [13,62]. Missing-data techniques using IBM like masks have been discussed in detail in previous chapters (see Chapters 14 and 15). Apart from missing-data ASR, other strategies have been proposed that use the IBM to improve ASR results. We will look at a few of them later in this chapter.

(v) Recently, Wang *et al*. [80] showed that IBM-modulated noise can produce intelligible speech. In this experiment, speech-shaped noise (SSN) is modulated by the IBM created for a mixture of speech and SSN. Speech shaped noise is broadband, and has a long-term spectrum matching that of natural speech. Even with a coarse frequency resolution (e.g., 16 bands), they observe nearly perfect intelligibility of IBM modulated noise.

Figure 16.2 shows an example of the IBM created for a two-talker mixture. The time-frequency representation used in the figure is called a *cochleagram*, which is commonly used in CASA [79]. Compared to the mixture in the middle left panel, the IBM-masked mixture (shown in the bottom left panel) is more similar to the target utterance (shown in the top left panel).

Apart from the IBM, research has also aimed at estimating the ideal ratio mask [3,73]. Note that, the real values in a ratio (soft) mask can also be interpreted as the probability of a T-F unit being target dominant. One can argue that estimating a ratio mask is computationally harder

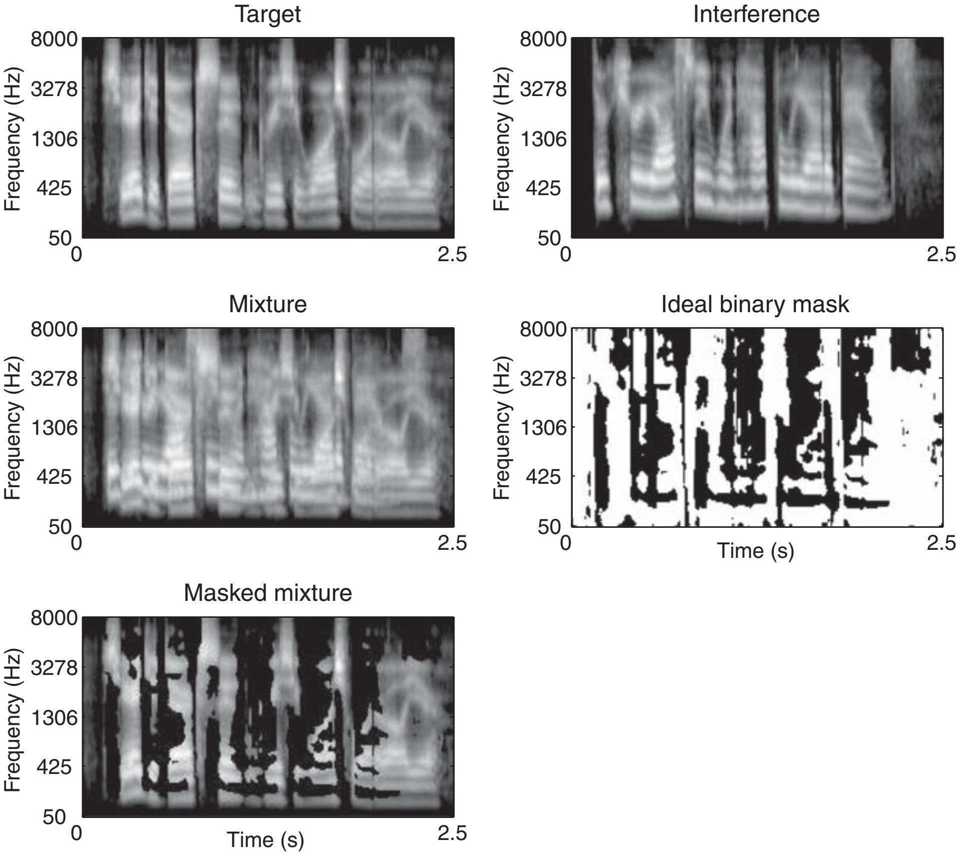

Figure 16.2 Illustration of the IBM. The top left panel shows a cochleagram of a target utterance where brightness indicates energy. The top right panel shows a cochleagram of the interference signal. The middle left panel shows a cochleagram of the mixture. The middle right panel shows the ideal binary mask for the mixture where a white pixel indicates 1 and a black pixel 0. The bottom left panel shows the cochleagram of the IBM-masked mixture.

than estimating a binary mask [77]. Nevertheless, the use of ratio masks has been shown to be advantageous in some ASR studies [3,73].

16.3.2 Typical CASA Architecture

Figure 16.3 shows a typical architecture of CASA. All CASA systems start with a peripheral analysis of the acoustic input (the mixture). Typically, the peripheral analysis converts the signal into a time-frequency representation. This is usually accomplished by using an auditory filter bank. The most commonly used is the gammatone filter bank [58]. The center frequencies of the gammatone filter bank are uniformly distributed on the ERB-rate scale [18]. ERB refers to the equivalent rectangular bandwidth of an auditory filter, which corresponds to the bandwidth

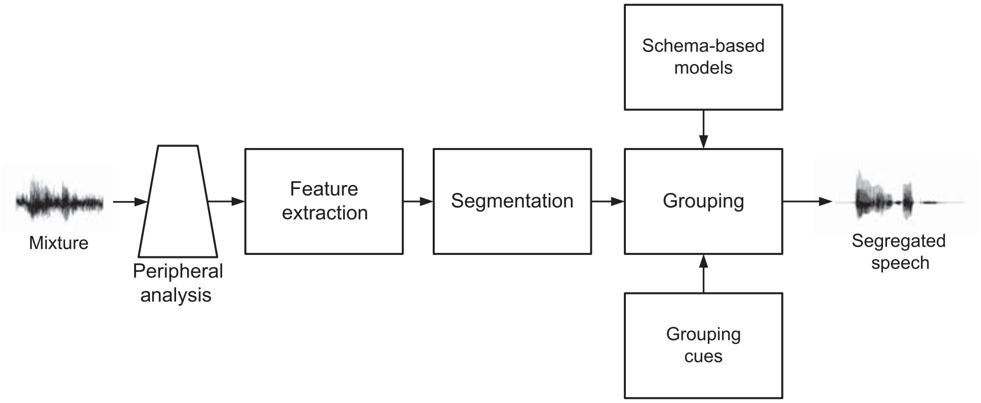

Figure 16.3 Schematic diagram of a typical CASA system.

of an ideal rectangular filter that has the same peak gain as the auditory filter with the same center frequency and passes the same total power for white noise. Similar to the Bark scale, the ERB-rate scale is a warped frequency scale akin to that of human cochlear filtering. The ERB scale is close to linear at low frequencies, but logarithmic at high frequencies. Figure 16.4 shows the responses of eight such filters, uniformly distributed according to the ERB-rate scale from 100 to 2000 Hz. Although eight filters are sufficient to fully span a frequency range of 50–8000 Hz, more filters (32 or 64) are typically used for a better frequency resolution. To simulate the firing activity of auditory nerve fibers, the output from the gammatone filter bank is further subjected to some nonlinear processing, where the Meddis hair cell model is typically used [48]. It models the rectification, compression and the firing pattern of the

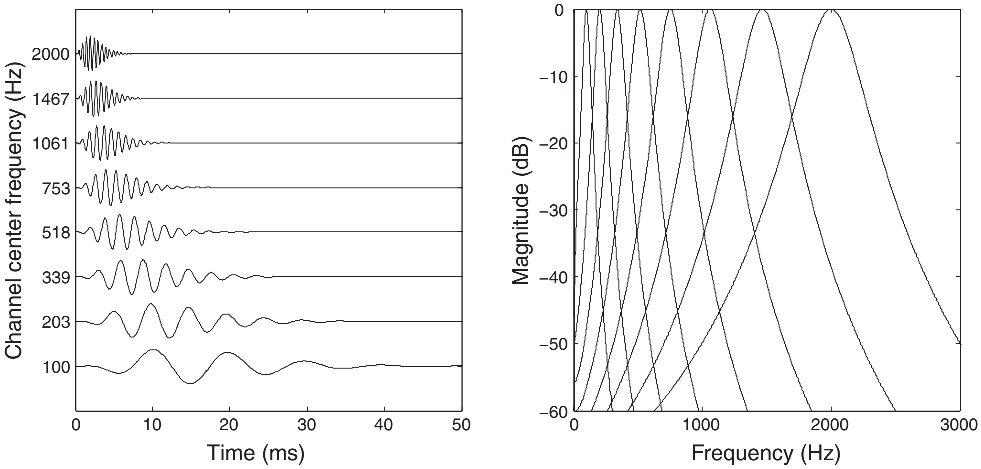

Figure 16.4 A gammatone filter bank. The left panel shows impulse responses of eight gammatone filters, with center frequencies equally spaced between 100 Hz and 2 KHz on the ERB-rate scale. The right panel shows the corresponding magnitude responses of the filters.

auditory nerve. Alternatively, a simple half wave rectification followed by some compression (square root or cubic root) can be used to model the nonlinearity. Finally, the output at each channel is windowed or downsampled. The result is the cochleagram of the acoustic signal as it models the processing performed by the cochlea [79]. An element of a cochleagram is a T-F unit, which represents the response of a particular filter at a time frame.

The next few stages vary depending on the specifics of different CASA systems. The feature extraction stage computes features such as $F0$, onset/offset, amplitude and frequency modulation. The extracted features enable the system to form segments, each of which is a contiguous region of T-F units. Segments provide a mid-level representation on which grouping operates. The grouping stage utilizes primitive and schema-based grouping cues. The output of the grouping stage can be an estimated binary mask or a ratio mask. Efficient algorithms exist that can resynthesize the target signal using a T-F mask and the original mixture signal [79,82].

16.4 CASA Strategies

Given the goal of estimating the IBM, we now discuss strategies to achieve it. The main focus of this section will be on monaural CASA techniques which have seen most of the development.

Monaural source segregation uses a single recording of the acoustic scene from which the target is to be segregated. The most important cue utilized for this task is the fundamental frequency. F0 estimation from clean speech is fairly accurate and many systems exist that perform well; for example Praat is a freely available tool which is widely used [6]. The presence of multiple sound sources in a scene adds to the complexity of the task as a single frame may now have multiple pitch points. Perhaps the earliest system that used F0 for speech segregation was proposed by Parsons [57]. He used the short-term magnitude spectrum of noisy speech to estimate multiple F0s. A sub-harmonic histogram method, proposed by Shroeder [64], was used to estimate the most dominant $F0$ in a frame. He then removed the harmonics of the estimated $F0$ from the mixture spectrum and used the remainder to estimate the second $F0$. The estimated $F0$s were finally used to segregate the mixture.

We start our discussion on IBM estimation in Section 16.4.1 by introducing strategies based on noise-estimation techniques from the speech-enhancement literature. More recent CASA-based strategies aim to segregate the target by extracting ASA cues like $F0$, amplitude modulation and onset/offset, which are then used to estimate the IBM. An alternative approach is to treat mask estimation as a binary classification problem. We explain these approaches in the subsequent subsections by treating two recent strategies in detail. The second subsection focuses on the *tandem* algorithm proposed by Hu and Wang [26] that uses several ASA cues to estimate the IBM. Section 16.4.3 focuses on a binary classification-based approach proposed by Kim *et al*. [36]. The final subsection briefly touches upon binaural CASA strategies.

16.4.1 IBM Estimation Based on Local SNR Estimates

In this sub-section, we discuss mask estimation strategies that are based on local signal-to-noise ratio estimates at each time-frequency unit. Such techniques typically make use of an estimate of the short-time noise power spectrum. The estimated noise power can be used to obtain the SNR and in turn a T-F mask. It should be clear from Equation (16.1) that with the true local SNR information, the IBM can be readily calculated. The noise estimate can also be used to define masks based on alternative criteria, like the negative energy criterion used by El-Maliki and Drygajlo [17]. We will first review a few noise-estimation techniques, followed by a brief discussion on how they can be used to estimate the IBM.

Noise (and SNR) estimation is a widely studied topic in speech enhancement largely in the context of spectral subtraction [5]. One commonly used technique is to assume that noise remains stationary throughout the duration of an utterance and that the first few frames are 'noise-only'. A noise estimate is then obtained by simply averaging the spectral energy of these frames. Such estimates are, for instance, used in Vizinho *et al.* [75], Josifovski *et al.* [34], Cooke *et al.* [13]. But noise is often nonstationary and therefore, such methods often result in poor IBM estimates. More sophisticated techniques have been proposed to estimate noise in nonstationary conditions. See, for example, voice-activity detection (VAD) [69] based methods [40], Hirsch's histogram based methods [23], recursive noise-estimation techniques [23], etc. Seltzer *et al.* [65] use an approach similar to Hirsch's to estimate the noise floor in each sub-band, which is in turn used for mask estimation (see Section 16.4.3). A more detailed discussion on noise estimation can be found in Chapter 4.

All noise-estimation techniques can be easily extended to estimate the SNR at each T-F unit by using it to obtain an estimate of the clean speech power spectrum. A spectral subtraction based approach [5,7] is commonly used, wherein the speech power is obtained by subtracting the noise power from the observed noisy spectral power. Further, a spectral floor is set and any estimate lower than the floor is automatically rounded to this preset value. Other direct SNR-estimation techniques have also been proposed in the literature. For example, Nemer *et al.* [53] utilize higher order statistics of speech and noise to estimate the local SNR, assuming a sinusoidal model for band restricted speech and a Gaussian model for noise. A supervised SNR-estimation technique was proposed by Tchorz and Kollmeier [74]. They use features inspired from psychoacoustics and a multilayer perceptron (MLP)-based classifier to estimate the SNR at each T-F unit. Interested readers are also referred to Loizou[46] for detailed reviews on these topics.

If a noise estimate is used to calculate the SNR, the IBM can be estimated using Equation (16.1) after setting the *LC* to an appropriate value. Although 0 dB is a natural choice here, other values have also been used [13,60]. Soft (ratio) masks can be obtained from local SNR estimates by applying a sigmoid function that maps it to a real number in the range $[0, 1]$, thereby allowing it to be interpreted as probability measures for subsequent processing. One can also define masks based on *a posteriori* SNR, which is the ratio of the noisy signal power to noise power expressed in *dB* [61]. This circumvents the need to estimate the clean speech power and local SNR. Note that any *a posteriori* SNR criterion can be equivalently expressed using a local SNR criterion. An even simpler alternative is to use the negative energy criterion proposed by El-Maliki and Drygajlo [17]. They identify reliable speech dominant units as those T-F units for which the observed noisy spectral energy is greater than the noise estimate. In other words, T-F units for which the spectral energy after subtracting the noise estimate from the observed noisy spectral energy is negative are considered noise dominant and unreliable. Raj and Stern [59] note that a combination of an SNR criterion and a negative energy criterion usually yields better quality masks.

In practice, such noise-estimation-based techniques work well in stationary conditions but tend to produce poor results in nonstationary conditions. Nonetheless, SNR-based techniques are still used because of their simplicity.

16.4.2 IBM Estimation using ASA Cues

The tandem system by Hu and Wang [26] aims at voiced speech segregation and F0 estimation in an iterative fashion. In describing the algorithm, we will explain how some of the ASA cues can be extracted and utilized for computing binary masks.

The tandem system uses several auditory representations that are widely used for pitch estimation. These representations are based on autocorrelation, which was originally proposed by Licklider back in the 1950s to explain pitch perception [44]. Autocorrelation has been used by other $F0$ estimation techniques [24,38,85]. The tandem system first uses a gammatone filter bank to decompose the signal into 128 frequency channels with center frequencies spaced uniformly in the ERB-rate scale from 50 to 8000 Hz. The output at each channel is divided into frames of length 20 ms with 10 ms overlap. A running autocorrelation function (ACF) is then calculated according to Equation (16.2) at each frame to form a *correlogram*:

$$A(t, f, \tau) = \frac{\sum_n x(tT_t - nT_n, f)x(tT_t - nT_n - \tau T_n, f)}{\sqrt{\sum_n x^2(tT_t - nT_n, f)}\sqrt{\sum_n x^2(tT_t - nT_n - \tau T_n, f)}}. \tag{16.2}$$

Here, $A(t, f, \tau)$ denotes the normalized autocorrelation function at frequency channel f and time frame *t*, and τ is the time delay in samples indexed by *n*. $T_t = 10$ ms and $T_n = 1/f_s$, where f_s is the sampling frequency, are the frame shift and the sampling time, respectively. The function is normalized so that the peak value at $\tau = 0$ is 1. An example of a correlogram is shown in Figure 16.5. Usually, a peak in the ACF corresponds to the time delay that represents a period of the signal. Since the target signal is speech, τ can be limited to the typical pitch range between 70 and 400 Hz, or τT_n between 2.5 and 15 ms [54]. Calculating the channel-wise ACF after decomposing the signal using a filter bank, instead of directly calculating it from the time domain signal, adds to the robustness of the F0 estimation process [14,85]. Additionally, a summary autocorrelation function (SACF) can be calculated by summing the ACFs across all the channels:

$$SACF(T, \tau) = \sum_f A(T, f, \tau). \tag{16.3}$$

A peak in the SACF corresponds to the time period that has support from many frequency channels. Since a periodic signal triggers responses in multiple channels, this peak likely indicates the period of the signal.

The cross-channel correlation between neighboring channels has been used to identify whether neighboring T-F units are dominated by the same source which can be used to group the units to form a segment [9,78]. Normalized cross-channel correlation, $C(t, f)$, is calculated

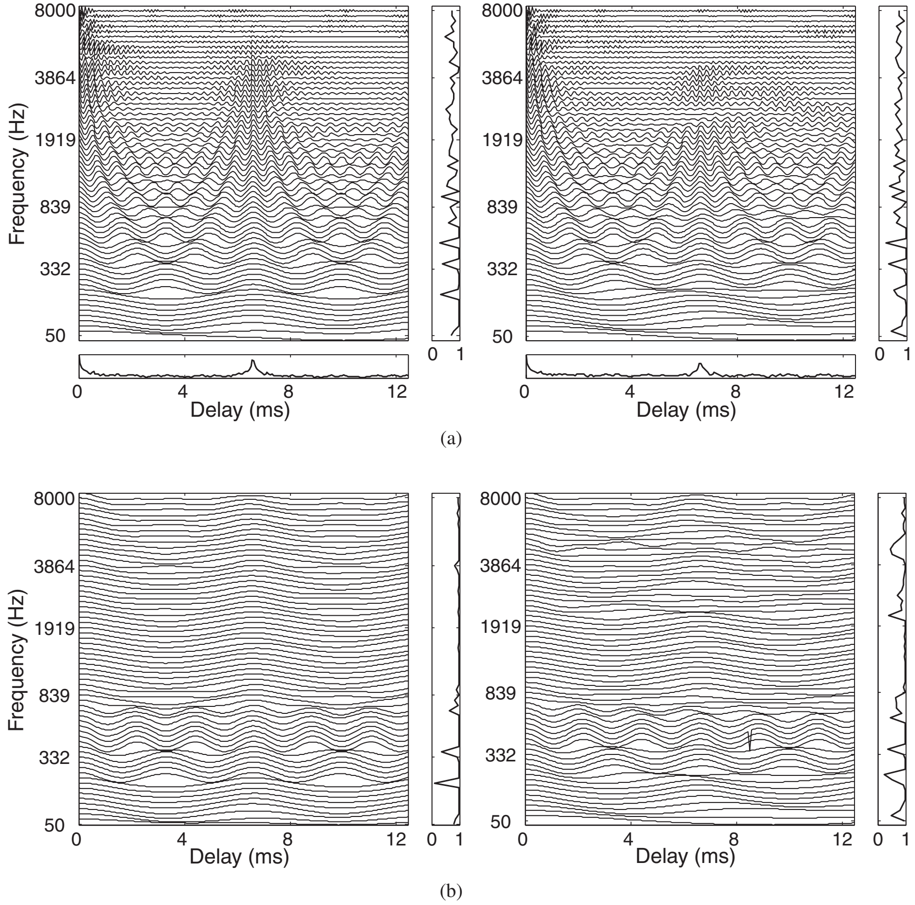

Figure 16.5 Autocorrelation and cross-channel correlation. (a) Correlogram at a frame for clean speech (top left panel) and a mixture of speech with babble noise at 6 dB SNR (top right panel). The corresponding cross-channel correlation and summary autocorrelation are shown on the right and the bottom panel of each figure, respectively. A peak in the SACF is clearly visible in both cases. Note that correlations of different frequency channels are represented using separate lines. (b) Corresponding envelope correlogram and envelope cross-channel correlation for clean speech (bottom left panel) and the mixture (bottom right panel). It can be clearly seen that the functions estimated from clean speech and noisy speech match closely.

using the ACF as:

$$C(t,f) = \frac{\sum_{\tau}\left[A(t,f,\tau)-\overline{A(t,f)}\right]\left[A(t,f+1,\tau)-\overline{A(t,f+1)}\right]}{\sqrt{\sum_{\tau}\left[A(t,f,\tau)-\overline{A(t,f)}\right]^2}\sqrt{\sum_{\tau}\left[A(t,f+1,\tau)-\overline{A(t,f+1)}\right]^2}}. \tag{16.4}$$

Here, $\overline{A(t,f)}$ denotes the mean of the ACF function over τ.

As mentioned earlier, gammatone filters with higher center frequencies have wider bandwidths (see Figure 16.4). As a result, for a periodic signal, high-frequency filters will respond to more than one harmonic of the signal. These harmonics are referred to as *unresolved*. Unresolved harmonics cause filter responses to be amplitude modulated, and the envelope of a filter response fluctuates at the fundamental frequency of the signal. This property has also been used as a cue to group segments and units in high-frequency channels [24]. Amplitude modulation or envelope can be captured by half wave rectification followed by band-pass filtering of the response. The pass band of the filter corresponds to the plausible pitch range of the signal. Replacing the filter responses in Equation (16.2) and Equation (16.4) with the extracted envelopes yields the normalized envelope autocorrelation, $A_{\mathrm{E}}(t,f,\tau)$, and the envelope cross-channel correlation, $C_{\mathrm{E}}(t,f)$, respectively. A_{E} can be used to estimate the periodicity of amplitude fluctuation. C_{E} encodes the similarity of the response envelopes of neighboring channels and aids segmentation. Figure 16.5 shows an example of a correlogram and an envelope correlogram for a single frame of speech (clean and noisy), and their corresponding cross-channel correlations and SACFs.

For T-F unit labeling, the tandem algorithm uses the probability that the signal within a unit is in agreement with a pitch period τ. This probability, denoted as $P(T,f,\tau)$, is estimated with the help of an MLP using a six-dimensional (6-D) pitch-based feature vector:

$$\begin{aligned} r(t,f,\tau) = [&A(t,f,\tau), \bar{f}(t,f)\tau - \mathrm{int}(\bar{f}(t,f)\tau), \mathrm{int}(\bar{f}(t,f)\tau), \\ &A_{\mathrm{E}}(t,f,\tau), \bar{f}_{\mathrm{E}}(t,f)\tau - \mathrm{int}(\bar{f}_{\mathrm{E}}(t,f)\tau), \mathrm{int}(\bar{f}_{\mathrm{E}}(t,f)\tau)], \end{aligned} \tag{16.5}$$

where the vector consists of ACFs and features derived using an estimate of the average instantaneous frequency, $\bar{f}(t,f)$. In the equation, int(.) returns the nearest integer and the subscript 'E' denotes envelope. $\bar{f}_{\mathrm{E}}$ is the instantaneous frequency estimated from the response envelope. If a signal is harmonically related to the pitch period τ, then int$(\bar{f}(t,f)\tau)$ and int$(\bar{f}_{\mathrm{E}}(t,f)\tau)$ will indicate a harmonic number. The difference between these products and their nearest integers in the second and the fourth terms quantifies a degree of this relationship. An MLP is trained for each filter channel in order to estimate $P(t,f,\tau)$[1].

The algorithm first estimates initial *pitch contours*, each of which is a set of contiguous pitch periods belonging to the same source, and their associated binary masks for up to two sound sources. The main part of the algorithm iteratively refines the initial estimates. The final stage applies onset/offset analysis to further improve segregation results. Let us now look at these stages in detail.

The initial stage starts by identifying T-F units corresponding to periodic signals. Such units tend to have high cross-channel correlation or envelope cross-channel correlation and, therefore, are identified by comparing $C(t,f)$ and $C_{\mathrm{E}}(t,f)$ with a threshold. Within each frame, the algorithm considers up to two dominant voiced sound sources. The identified T-F units of a frame are grouped using a two step process. First, estimate up to two $F0$s. Next, assign a T-F unit to an $F0$ group if it agrees with the $F0$.

An earlier model by Hu and Wang [24] identifies the dominant pitch period of a frame as the lag that corresponds to the maximum in the summary autocorrelation function (Equation (16.3)). To check if a T-F unit agrees with the dominant pitch period, they compare the value

[1] Note that this term is a convenient abuse of notation. It, in fact, represents the posterior probability of the T-F unit being in agreement with the pitch period given the 6-D pitch-based features.

of the ACF at the estimated pitch period to the peak value of the ACF for that T-F unit:

$$\frac{A(t, f, \tau_D(t))}{A(t, f, \tau_P(t, f))} > \theta_P. \tag{16.6}$$

Here, $\tau_D(t)$ and $\tau_P(t, f)$ are the delays that correspond to the estimated $F0$ and the maximum in the ACF, respectively, for channel *f* at time frame *t*. If the signal within the T-F unit has a period close to the estimated $F0$, then this ratio will be close to 1. θ_P defines a threshold to make a binary decision about the agreement.

The tandem algorithm uses a similar approach, but instead of the ACF it uses the probability function, $P(t, f, \tau)$, estimated using the MLPs. Having identified the T-F units of each frame with strong periodicity, the algorithm chooses the lag, τ, that has the most support from these units as the dominant pitch period of the frame. A T-F unit is said to support τ if the probability, $P(t, f, \tau)$, is above a chosen threshold. The T-F units that support the dominant pitch period are then grouped together. The second pitch period and the associated set of T-F units are estimated in a similar fashion, using those units not in the first group. To remove spurious pitch estimates, if there are too few supporting T-F units, the estimated pitch is discarded.

To form pitch contours from these initial estimates, the algorithm groups the pitch periods of any three consecutive frames if their values change by less than 20% from one frame to the next. The temporal continuity of the sets of T-F units associated with the pitch periods is also considered before grouping pitch estimates together; at least half of the frequency channels associated with the pitch periods of neighboring frames should match for them to be grouped into a pitch contour. After the initial stage, each pitch contour has an associated T-F mask. Since pitch changes rather smoothly in natural speech, each of the formed pitch contours and its associated binary mask usually belong to a single sound source. Isolated pitch points after this initial grouping are considered unreliable and discarded.

These initial estimates are then refined using an iterative procedure. The idea is to use obtained binary masks to obtain better pitch contours, and then use the refined pitch estimates to re-estimate the masks. Each iteration of the tandem algorithm consists of two steps:

(i) The first step expands each pitch contour to its neighboring frames, and re-estimates its pitch periods. Since pitch changes smoothly over time, the pitch periods of the contour can be used to estimate potential pitch periods in the contour's neighboring frames. Specifically, for the *k*th pitch contour τ_k, that extends from frame t_1 to t_2, the corresponding binary mask, $M_k(t)$ $(t = t_1, \ldots, t_2)$, is extended to frames $t_1 - 1$ and $t_2 + 1$ by setting $M_k(t_1 - 1) = M_k(t_1)$ and $M_k(t_2 + 1) = M_k(t_2)$. Using this new mask, the periods of the pitch contour are reestimated. A summary probability function, $SP(t, \tau)$, which is similar to $SACF(t, \tau)$ but uses $P(t, f, \tau)$ values instead of $A(t, f, \tau)$, is calculated at each frame for this purpose. The *SP* function tends to have significant peaks at multiples of the pitch period. Therefore, an MLP is trained to choose the correct pitch period from among the multiple candidates. The expansion stops at either end when the estimated pitch violates temporal continuity with the existing pitch contour. Note that, as a result of contour expansion, pitch contours may be combined.

(ii) The second step reestimates the mask corresponding to each of the pitch contours. This is done by identifying T-F units of each frame that are in agreement with the estimated pitch period of that frame. Given the pitch period $\tau_D(t)$, $P(t, f, \tau_D)$ can be directly used to make this decision at each T-F unit. But this does not take into consideration the temporal continuity and the wide-band nature of speech. If a T-F unit is in agreement with τ_D, its neighboring T-F units also tend to agree with τ_D. For added robustness, the tandem algorithm trains an MLP to perform unit labeling based on a neighboring set of T-F units. It takes as input the $P(t, f, \tau)$ values of a set of neighboring T-F units, centered at the unit for which the labeling decision has to be made. The output of this MLP is finally used to label each T-F unit.

The algorithm iterates between these two steps until it converges or the number of iterations exceeds a predefined maximum (20 is suggested).

The final step of the tandem algorithm is a segmentation stage based on onset/offset analysis, which may be viewed as post processing. The stage forms segments by detecting sudden changes in intensity as such a change indicates an onset or offset of an acoustic event. As discussed earlier, onset and offset are prominent ASA principles (see Figure 16.1). Segments are formed using multiscale analysis of onsets and offsets (see Hu and Wang [25] for details). The tandem algorithm further breaks each segment down to channel wise subsegments, called *T-segments* as they span multiple time frames but are restricted to a single frequency channel. Each T-segment is then classified as a whole as target dominant if at least half its energy is contained in the voiced frames of the target and at least half of the energy in these voiced frames is included in the target mask. If the conditions are not satisfied, the labeling from the iterative stage remains unchanged for the units of the T-segment.

Figure 16.6 illustrates the results of different stages of the tandem system. The mask obtained at the end of the iterative stage (Figure 16.6(e)) includes most of the target speech. The subsequent segmentation stage improves the segregation results by recovering a few previously masked (mask value 0) T-F units, for example toward the end of the utterance in Figure 16.6(g). These units were identified from the onset/offset segments. The final resynthesized waveform, shown in Figure 16.6(h), is close to the original signal (Figure 16.6(b)).

There are two important aspects of CASA that the tandem algorithm does not consider. The first one is sequential organization. The outputs of the tandem system are multiple pitch contours and associated binary masks. The pitch track (and therefore the mask) of a target utterance need not be continuous as there are breaks due to silence and unvoiced speech. Sections before and after such discontinuities have to be sequentially grouped into the target stream. The tandem system assumes ideal sequential grouping, and therefore ignores the sequential grouping issue. Methods for sequential grouping have been proposed. Barker *et al.* [4] proposed a schema based approach using ASR models to simultaneously perform sequential integration and speech recognition (more about this in Section 14.4.3). Ma *et al.* [47] later used a similar approach to group segments that were formed using correlograms in voiced intervals and a watershed algorithm in unvoiced intervals. Shao and Wang [67] proposed a speaker model-based approach for sequential grouping. Recently, Hu and Wang [30] proposed an unsupervised grouping strategy based on clustering and reported results comparable to the model-based approach of Shao and Wang.

The second issue with the tandem algorithm is that it does not deal with unvoiced speech. An analysis by Hu and Wang [28] shows that unvoiced speech accounts for more than 20% of spoken English, measured in terms of both frequency and duration of speech sounds. Therefore, unvoiced speech segregation is important for improving the intelligibility and ASR of the segregated target signal. Dealing with unvoiced speech is challenging as it has noise-like characteristics and lacks strong grouping cues such as $F0$. Hu and Wang [28]

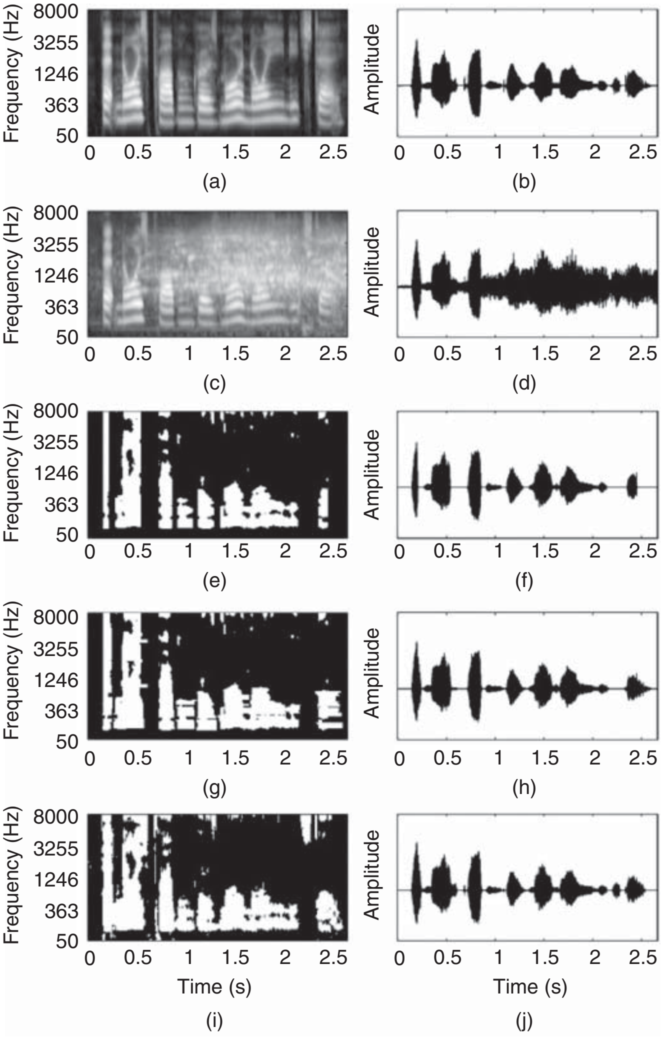

Figure 16.6 Different stages of IBM estimation using the tandem system. (a) Cochleagram of a female target utterance. (b) Corresponding waveform. (c) Cochleagram of a mixture signal obtained by adding crowd noise to the target utterance. (d) Corresponding waveform. (e) Mask obtained at the end of the iterative stage of the algorithm. (f) Waveform of the resynthesized target using the mask. (g) The final mask obtained after the segmentation stage. (h) The resynthesized waveform. (i) The IBM. (j) Resynthesized signal using the IBM. Reproduced by permission from Hu and Wang [26] © 2010 IEEE.

suggest a method to extract unvoiced speech using onset/offset based segments. They first segregate voiced speech. Then, acoustic-phonetic features are used to classify the remaining segments as interference dominant or unvoiced speech dominant. A simpler system was later proposed by Hu and Wang [29]. Their system first segregates voiced speech and removes other periodic intrusions from the mixture. It then uses a spectral subtraction based scheme to obtain segments in unvoiced intervals (an unvoiced interval corresponds to a contiguous group of unvoiced frames); the noise estimate for each unvoiced interval is estimated using the mixture energy in the masked T-F units of its neighboring voiced intervals. Together with an approximation of the target energy obtained by subtracting the estimated noise from the mixture, the local SNR at each T-F unit is calculated. The segments themselves are formed by grouping together neighboring T-F units that have estimated SNRs above a chosen threshold. The obtained segments are then classified as target or interference dominant based on the observation that most of the target dominant unvoiced speech segments reside in the high-frequency region. The algorithm works well if the noise remains fairly stationary during the duration of an unvoiced interval and the neighboring voiced intervals.

16.4.3 IBM Estimation as Binary Classification

The tandem algorithm exemplifies a system that uses ASA cues and supervised learning to estimate the IBM. When it comes to direct classification, the issues lie in choosing appropriate features that can discriminate target speech from interference, and an appropriate classifier. To explain how direct classification is applied, we describe the classification-based approach of Kim *et al*. [36] in detail.

The system by Kim *et al*. uses amplitude modulation spectrograms (AMS) as the feature to build their classifier. To obtain AMS features, the signal is first passed through a 25 channel filter bank, with filter center frequencies spaced according to the mel-frequency scale. The output at each channel is full-wave rectified and decimated by a factor of 3 to obtain the envelope of the response. Next, the envelope is divided into frames 32 ms long with 16 ms overlap. The modulation spectrum at each T-F unit is then calculated using the FFT[2]. The FFT magnitudes are finally integrated using 15 triangular windows spaced uniformly from 15.6 to 400 Hz, resulting in 15 AMS features [39]. Kim *et al*. augment the extracted AMS features with delta features calculated from the neighboring T-F units. The delta features are calculated across time and frequency, and for each of the 15 features separately. They help capture temporal and spectral correlations between T-F units. This creates a 45-dimensional feature representation for each T-F unit, $AMS(t, f)$.

Given the 45-dimensional input, a Gaussian mixture model (GMM)-based classifier is trained to do the classification. The desired unit labels are set using the IBM created using an *LC* (see Equation (16.1)) of –8 dB for low-frequency channels (channels 1 through 15) and –16 dB for high-frequency channels (channels 16 through 25). This creates a group of masked T-F units, represented as λ_0, and unmasked (mask value 1) T-F units, λ_1. The authors chose a lower *LC* for high-frequency channels to account for the difference in the masking characteristics of speech across spectrum. Each group, λ_i, where $i = 0, 1$, is further divided into two smaller subgroups, λ_i^0 and λ_i^1, using a second threshold, LC_i. The thresholds ($LC_0 < LC$ and $LC_1 > LC$) are chosen such that the amount of training data in the two subgroups of a group are the same. This second subdivision is done mainly to reduce the training time of the GMMs. A 256-mixture, 45-dimensional, full-covariance GMM is trained using the expectation-maximization algorithm to model the distribution of each of the 4 subgroups.

[2] A T-F unit, here, refers to a 32 ms long frame at a particular frequency channel.

Given a T-F unit from a noisy utterance, a Bayesian decision is then made to obtain a binary label that is 0 if and only if $P(\lambda_0 \mid AMS(t,f)) > P(\lambda_1 \mid AMS(t,f))$, where

$$\begin{aligned} P(\lambda_0 \mid AMS(t,f)) &= \frac{P(\lambda_0, AMS(t,f))}{P(AMS(t,f))} \\ &= \frac{P(\lambda_0^0)P(AMS(t,f) \mid \lambda_0^0) + P(\lambda_0^1)P(AMS(t,f) \mid \lambda_0^1)}{P(AMS(t,f))}. \end{aligned}$$

The equation calculates the *a posteriori* probability of λ_0 given the AMS features at the T-F unit. $P(\lambda_0^0)$ and $P(\lambda_0^1)$ are the *a priori* probabilities of subgroups λ_0^0 and λ_0^1, respectively, calculated from the training set. The likelihoods, $P(AMS(t,f) \mid \lambda_0^0)$ and $P(AMS(t,f) \mid \lambda_0^1)$, are estimated using the trained GMMs. $P(AMS(t,f))$ is independent of the class label and, hence, can be ignored. $P(\lambda_1 \mid AMS(t,f))$ is calculated in a similar fashion.

One advantage of using the AMS feature is that it can handle both voiced and unvoiced speech, as opposed to the 6-D pitch based feature used by the tandem algorithm which can be used only to classify voiced speech. As a result, the mask obtained using Kim *et al.*'s algorithm includes both voiced and unvoiced speech.

Figure 16.7 shows an estimated binary mask using Kim *et al.*'s algorithm. The authors evaluated their system using speech intelligibility tests and reported substantial improvements in the intelligibility of segregated speech for normal-hearing listeners [36]. It is worth emphasizing that this is the first monaural segregation system that produces improved speech intelligibility.

One of the main disadvantages of Kim *et al.*'s system is that training is noise dependent. Although it works well when tested on speech corrupted with the same noise types, the performance degrades significantly when previously unseen noise types are used during the testing stage. A second disadvantage of the system is that it can handle only nonspeech intrusions because AMS features mainly distinguish speech and nonspeech signals. By avoiding competing talkers, the problem of sequential organization is avoided because all detected speech belongs to the target.

Jin and Wang [32] also proposed a classification-based approach to perform voiced speech segregation in reverberant environments. For T-F unit classification, they use the 6-D pitch-based features given in Equation (16.5), and an MLP-based classifier. In order to utilize global information that is not sufficiently represented at the T-F unit level, an additional segmentation stage is used by their system. Segmentation is performed based on cross-channel correlation and temporal continuity in low-frequency channels—adjacent T-F units with high cross-channel correlation are iteratively merged to form larger segments. In high-frequency channels, they are formed based on onset/offset analysis [25]. The unit level decisions are then used to group the formed voiced segments either with the target stream or the nontarget (or the background) stream. Their system produced good segregation results under various reverberant conditions. Since pitch-based features are derived using the pitch of the *target*, classifiers trained on such features tend to generalize better than those trained using AMS features.

More recently, Kun and Wang proposed an SVM-based binary mask estimation model [19]. Inspired by Jin and Wang [32] and Kim *et al.* [36], they propose to combine pitch-based and AMS features along with the use of an SVM based classifier. Their system performs well in a variety of test conditions and is found to have good generalization to unseen noise types.

In the context of robust ASR, Seltzer *et al.* [65] proposed a similar Bayesian classification based approach to mask estimation. They extract the following features at the T-F unit level to

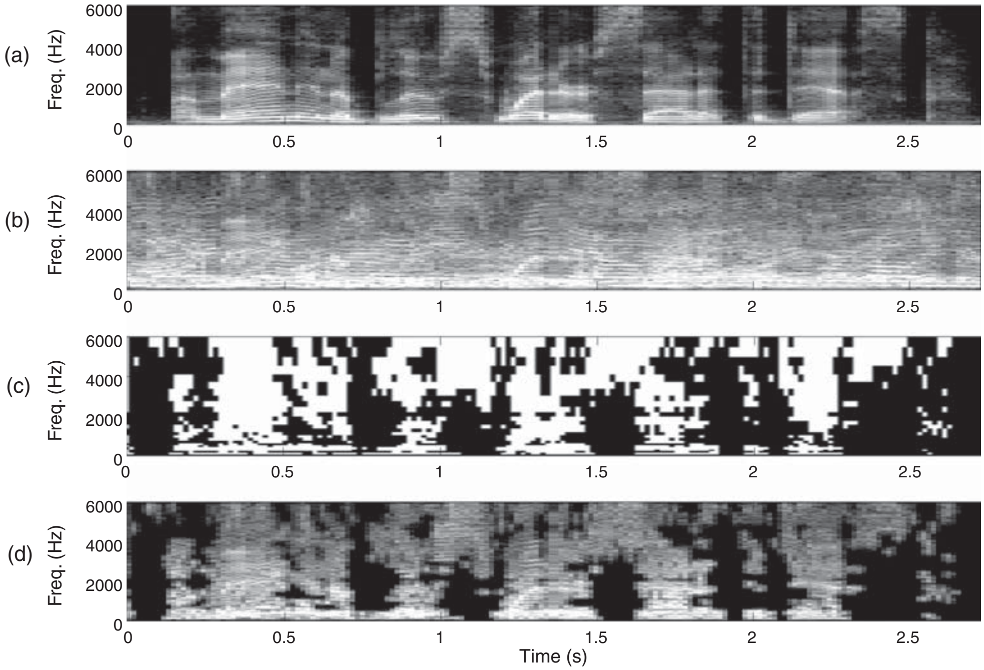

Figure 16.7 IBM estimation using classification. (a) A spectrogram of a target utterance from the IEEE corpus. (b) Spectrogram of the noisy mixture obtained by adding babble noise to the target utterance. (c) The estimated binary mask. (d) The spectrogram of the resynthesized signal obtained using the estimated binary mask. Reprinted with permission from Kim *et al*. [36] © 2009, Acoustical Society of America.

build GMM based Bayesian classifiers: comb filter ratio (CFR), which is the log ratio of the total energy at the harmonics of the fundamental frequency estimated for a frame to the total energy in between those frequencies; autocorrelation peak ratio (APR), which is the ratio of height of the largest secondary peak in the ACF to the height of the main peak; the log ratio of the energy within the T-F unit to the total energy at that time frame; kurtosis, calculated from sample averages in each subband at each time frame; spectral flatness, measured in terms of the variance of the subband energy within the spectrographic neighborhood of the T-F unit; the ratio of the subband energy at each time frame to the noise floor estimated for that subband; and spectral subtraction based local SNR estimate. The features are chosen such that they capture the characteristics of speech in noise without making assumptions about the underlying noise type. Except for the first two features, viz. CFR and APR, the remaining ones can be used to characterize properties of T-F units in both voiced and unvoiced time frames. CFR and APR are used only for the T-F units in voiced frames. GMMs are trained for voiced and unvoiced speech separately, and also at each subband and are in turn used to obtain soft T-F masks. The obtained masks improve ASR performance when used in conjunction with missing-data-based strategies. Seltzer *et al*. [65] use speech mixed with white noise to train the classifiers. This can be limiting when it comes to generalization to unseen noisy conditions. To overcome this, Kim and Stern [37] suggest training each frequency band separately, using artificial colored noise signals generated specifically for each band. They show that this can yield better generalization results as compared to using white noise alone for training.

In a way, classification-based strategies simplify the task of speech segregation, at least conceptually. It bypasses the steps of a typical CASA system which extracts perceptually motivated cues and applies the ASA stages of segmentation and grouping to obtain a binary mask. The potential downside of relying on supervised learning is the perennial issue of generalization to unseen conditions.

16.4.4 Binaural Mask Estimation Strategies

Binaural CASA systems use two microphone recordings to segregate the target from the mixture. Most binaural systems try to extract localization cues, for example azimuth, which are encoded in the differences between the signals that reach the two ears (or microphones). In this regard, interaural time difference (ITD) and interaural intensity difference (IID) are the two most important cues. ITD is the difference between the arrival times of the signal at the two ears. ITD is ambiguous at high frequencies ($>$ 1.5 KHz) because of short wavelengths as compared to the distance between the ears. IID is the difference in the intensity of the sound that reaches the two ears, usually expressed in decibels, and it occurs because of the 'shadow' effect of the human head. Contrary to ITD, IID is not useful at low frequencies ($<$ 500 Hz) because such low-frequency sound components diffract around the head overcoming the shadow effect in the process.

Two classical strategies strongly influenced binaural segregation: the cross-correlation based model for ITD estimation proposed by Jeffress [31] and the equalization-cancellation (EC) model of Durlarch [16]. The EC model tries to segregate the target in a two stage process. In the first stage, the noise levels in the signals arriving at the two ears are equalized. This is followed by subtraction of the signals at the two ears in the cancellation stage. The noise equalized in the first stage gets canceled during the second stage, producing a cleaner target. The Jeffress model is based on the similarity of the signals that arrive at the two ears. The neural firing patterns of the two ears are passed through delay lines; the delay that maximizes the correlation between the two patterns is identified as the ITD of the signal.

To compute ITD, a normalized cross-correlation function, $C(t, f, \tau)$, is typically used

$$C(t,f,\tau) = \frac{\sum\limits_{n} x_L(tT_t - nT_n, f)x_R(tT_t - nT_n - \tau T_n, f)}{\sqrt{\sum\limits_{n} x_L^2(tT_t - nT_n, f)}\sqrt{\sum\limits_{n} x_R^2(tT_t - nT_n - \tau T_n, f)}}. \tag{16.7}$$

The above equation calculates cross-correlation at frequency channel f and time frame t, for a time lag τ. x_L and x_R correspond to the left and right ear response, respectively. T_t and T_n have the same meanings as in Equation (16.2). Similar to the normalized autocorrelation function, the cross-correlation function will have a peak at a delay that relates to ITD. IID can be calculated as the ratio of the mean power of the signals that arrive at the two ears:

$$IID(t,f) = 10\log_{10}\left(\frac{\sum_n x_L^2(tT_t - nT_n, f)}{\sum_n x_R^2(tT_t - nT_n, f)}\right). \tag{16.8}$$

An IBM estimation strategy based on classifying ITD and IID estimates was proposed by Roman *et al*. [62], which is probably the first classification-based system for speech segregation. They observed that, given a predefined configuration of the target and the interference (configuration here refers to the azimuths of the target and the interference), ITD and IID values vary smoothly and systematically with respect to the relative strength of the target and the mixture. This prompted them to model the distribution of target dominant units and interference dominant units of each frequency channel in the ITD-IID space. Their system models the distributions using a nonparametric kernel-density estimator. For an unseen test utterance, the binary decision at each T-F unit is made by comparing the probabilities of the unit being target dominant and interference dominant, given the observed ITD and IID at that unit. The binary masks estimated by their model are very close to the IBM, with excellent performances in terms of SNR gains, speech intelligibility and ASR accuracies. The main drawback of the model is that ITD-IID distributions are configuration dependent. A similar system was proposed by Harding *et al*. [20], which assumes that only the target azimuth is known *a priori*. It then learns the joint distribution of ITD and IID for target dominant T-F units using a histogram-based method. These distributions are used to predict the probability of a unit being target dominant from the observed ITD and IID. The estimated probabilities are directly used in the form of a ratio mask, to improve ASR results in reverberant conditions.

The above strategies are based on modeling the distribution of the binaural cues in the ITD-IID space. An alternative approach was proposed by Palomaki *et al*. [55]. This approach first estimates target and interference azimuths. It then classifies a T-F unit as target or interference dominant by comparing the values of the cross-correlation function at the estimated azimuths of the target and the interference. In order to deal with room reverberation, their system models the precedence effect [45] by using the low-pass filtered envelope response of each channel as an inhibitor. This reduces the effect of late echoes in reverberant situations by preserving transient and suppressing sustained responses. Palomaki *et al*. reported good ASR results in reverberant situations using the above algorithm to estimate binary masks.

Recently, Woodruff and Wang [84] proposed a system that combines monaural and binaural cues to estimate the IBM. Their system uses a monaural CASA algorithm to first obtain simultaneous streams, each occupying a continuous time interval. They use the tandem algorithm, described earlier, for this purpose. Binaural cues are then used to jointly estimate the azimuths of the streams that comprise the scene and their corresponding sets of sequentially grouped simultaneous streams.

16.5 Integrating CASA with ASR

The CASA strategies discussed in Section 16.4 provide us several perceptually inspired ways of segregating the target from a mixture. The main focus has been on estimating the ideal binary mask. Although IBM-based strategies produce good segregation results, integrating CASA and ASR has not been as straightforward a task as it seems. A simple way of combining CASA with ASR is to use CASA as a preprocessor. ASR models trained in clean conditions can then be used to perform recognition on the segregated target speech. This can be problematic. Even when the IBM is used, the resynthesized signal will have artifacts that may pose challenges to recognition. Errors in IBM estimation will further degrade the performance of such systems.

Nevertheless, CASA has been used as a preprocessor in some systems and has been shown to produce good results. One such model was proposed by Srinivasan *et al*. [73]. Their system uses a ratio T-F mask to enhance a noisy utterance. A conventional HMM-based ASR system trained using the mel-frequency cepstral coefficients (MFCC) of clean speech is used to recognize the enhanced speech. For mask estimation, they use the binaural segregation model by Roman *et al*. [62]. Srinivasan *et al*. compared their system with the missing-data ASR approach [13] and found that using such a CASA-based preprocessor can be advantageous as the vocabulary size of the recognition task increases. The limitation of missing-data ASR in dealing with larger vocabulary tasks had been reported earlier [60]. The use of a ratio mask instead of a binary mask coupled with accurate mask estimation helped their system in overcoming some of the limitations of using CASA as a preprocessor.

More recently, Hartmann and Fosler-Lussier [21] compared the performance of an ASR system that simply discards masked T-F units, which is equivalent to processing the noisy speech with a binary mask, with a system that reconstructs those units based on the information available from the unmasked T-F units. Such feature-reconstruction strategies have been used to improve noise robust ASR [60]. An HMM based ASR system trained in clean conditions is used to perform recognition. They observe that the direct use of IBM-processed speech performs significantly better than the reconstructed speech, and yields ASR results only a few percentage points worse than those in clean conditions. When noise is added to the IBM by randomly flipping 1s and 0s, only after the amount of mask errors exceeds some point does reconstruction work better. This is a surprising observation, considering the conventional wisdom that the binary nature of a mask is supposed to skew the cepstral coefficients (they used PLP cepstral coefficients to build their ASR system). This study points to the need of a deeper understanding of the effects of using binary masks on ASR performance.

The above methods somehow modify the features so that they can be used with ASR models trained in clean conditions. Such strategies have been called *feature compensation* or *source-driven* methods. Feature compensation includes techniques that use CASA based strategies for segregating the target [21,73] and reconstructing unreliable features [60]. An alternative approach would be to modify ASR models so that they implicitly accommodate missing or corrupt speech features. Such strategies have been termed *model compensation* or *classifier compensation* methods. The missing-data ASR techniques are examples of model compensation strategies [13]. There are also strategies that combine feature compensation and model compensation [15,71], and simultaneously perform CASA and ASR [4,72].

A much simpler strategy for integrating CASA and ASR was proposed by Narayanan and Wang [50] and Karadogan *et al*. [35]. They interpret IBMs as binary images and use a binary pattern classifier to do ASR. The idea of using binary pattern recognition for ASR is radically different from the existing strategies that use detailed speech features like MFCCs. Their work was motivated by the speech perception study showing that modulating noise by the IBM can produce intelligible speech for humans [80, also see Section 16.3]. Since noise carries no speech information, intelligibility must be induced by the binary pattern of the IBM itself. This indicates that the pattern carries important phonetic information. The system described in Narayanan and Wang [50] is designed for an isolated digit recognition task. The ASR module is based on convolutional neural networks [41,68], which have previously been used successfully for handwritten digit and object recognition. Their system obtains reasonable results even when the IBM is estimated directly from noisy speech using a CASA algorithm. They extend their system further in Narayanan and Wang [51] to perform a more challenging phone classification task, and show that IBMs and traditional speech features like MFCCs carry complimentary information that can be combined to improve the overall classification performance. The combined system obtains classification accuracies that compare favorably to most of the results reported in recent phone classification literature. It is quite interesting to note that features that are based on *binary* patterns can obtain good results on complex ASR tasks. Such CASA inspired features may eventually be needed for achieving robust ASR.

In the following subsection we discuss in greater detail an example of a CASA-inspired ASR framework. The subsection focuses on the uncertainty transform model proposed by Srinivasan and Wang [71] that combines feature compensation and model compensation to improve ASR performance.

16.5.1 Uncertainty Transform Model

Using a speech-enhancement algorithm to obtain features for ASR does not always yield good recognition results. This is because, even with the best enhancement algorithms, the enhanced features remain somewhat noisy, as far as the ASR models trained in clean conditions are concerned. Moreover, the variance of such features, with respect to the corresponding clean features, varies across time and frequency. Uncertainty decoding has been suggested as a strategy to modify ASR model parameters to take into account the inherent uncertainty of such enhanced features (see Chapter 17 for a more detailed handling of uncertainty decoding strategies). It has been shown that feature uncertainties contribute to an increase in the variance of trained acoustic variables and accounting for it during the recognition (decoding) stage can significantly improve ASR performance [15].

A mismatch in the domain of operation between speech enhancement or segregation and ASR can pose problems in effectively adjusting ASR model parameters based on estimated uncertainty. Such a mismatch exists for most CASA-based techniques as they operate either in the spectral or T-F domain, as opposed to ASR models that operate in the cepstral domain. Training ASR models in the spectral domain is known to produce suboptimal performance. In order to overcome this mismatch problem, Srinivasan and Wang [71] suggested a technique to transform the uncertainties estimated in the spectral domain to the cepstral domain.

The uncertainty transform model by Srinivasan and Wang consists of a speech-enhancement module, an uncertainty transformer, and a traditional HMM-based ASR module that operates in the cepstral domain. The enhancement module uses a spectrogram reconstruction method that is similar to [60] but operates in the linear spectral domain. To perform recognition, the enhanced spectral features are transformed to the cepstral domain. The corresponding uncertainties, originally estimated in the spectral domain, are transformed using a supervised learning method. Given the enhanced cepstral features and associated uncertainties, recognition is performed in an uncertainty decoding framework. Details about these stages are discussed below.

The speech-enhancement module starts by converting a noisy speech signal into the spectral domain using the FFT. The noisy spectrogram is then processed using a speech-segregation algorithm that estimates the IBM. A binary mask partitions a noisy spectral vector, $\mathbf{y}$, into its reliable components, $\mathbf{y}_r$, and the unreliable components, $\mathbf{y}_u$. Assuming that $\mathbf{y}_r$ sufficiently approximates the corresponding clean speech spectral values, $\mathbf{x}_r$, the goal of reconstruction is to approximate the true spectral values, $\mathbf{x}_u$, of the unreliable components. It uses a speech prior model for this purpose, implemented as a large GMM, where the probability density of a spectral vector of speech ($\mathbf{x}$) is modeled as

$$p(\mathbf{x}) = \sum_{k=1}^{K} P(k)p(\mathbf{x} \mid k).$$

Here, *K* represents the number of Gaussians in the GMM, *k* is the Gaussian index, $P(k)$ is the prior probability of the *k*th component (or the component weight), and $p(\mathbf{x} \mid k) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Theta}_k)$ is the conditional probability density of $\mathbf{x}$ given the *k*th Gaussian. In the Gaussian, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Theta}_k$ denote the mean vector and the covariance matrix, respectively. Such a GMM can be trained by pooling the entire training data and using an expectation maximization algorithm to learn the parameters. The mean and the covariance matrix of the *k*th Gaussian are also partitioned into its reliable and unreliable components using a binary mask:

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_{r,k} \\ \boldsymbol{\mu}_{u,k} \end{bmatrix}, \boldsymbol{\Theta}_k = \begin{bmatrix} \boldsymbol{\Theta}_{rr,k} & \boldsymbol{\Theta}_{ru,k} \\ \boldsymbol{\Theta}_{ur,k} & \boldsymbol{\Theta}_{uu,k} \end{bmatrix},$$

where $\boldsymbol{\mu}_{r,k}$ and $\boldsymbol{\mu}_{u,k}$ are the reliable and the unreliable components of the mean vector of the *k*th Gaussian, respectively; $\boldsymbol{\Theta}_{rr,k}$ and $\boldsymbol{\Theta}_{uu,k}$ are the corresponding covariances of the reliable and the unreliable components; and $\boldsymbol{\Theta}_{ru,k}$ and $\boldsymbol{\Theta}_{ur,k}$ are the cross-covariances.

The unreliable components are reconstructed by first estimating the *a posteriori* probability of the *k*th Gaussian using only the reliable components, $\mathbf{x}_r$, of the frame:

$$P(k \mid \mathbf{x}_r) = \frac{P(k)p(\mathbf{x}_r \mid k)}{\sum_{k=1}^{K} P(k)p(\mathbf{x}_r \mid k)}. \tag{16.9}$$

Next, the conditional mean of the unreliable components given the reliable components is approximated as

$$\widehat{\boldsymbol{\mu}}_{u,k} = \boldsymbol{\mu}_{u,k} + \boldsymbol{\Theta}_{ur,k}\boldsymbol{\Theta}_{rr,k}^{-1}(\mathbf{x}_r - \boldsymbol{\mu}_{r,k}). \tag{16.10}$$

Note that this is the standard formula for calculating the conditional mean of random variables that follow a multivariate normal distribution.

Given the *a posteriori* component weights and the conditional mean, a good approximation of the unreliable components is the expected value of $\mathbf{x}_u$ given $\mathbf{x}_r$, which is also the minimum mean-squared estimate (MMSE) of $\mathbf{x}_u$. The MMSE estimate can be calculated as

$$\widehat{\mathbf{x}}_u = E_{\mathbf{x}_u \mid \mathbf{x}_r}(\mathbf{x}_u) = \sum_{k=1}^{K} P(k \mid \mathbf{x}_r)\widehat{\boldsymbol{\mu}}_{u,k} \tag{16.11}$$

Finally, a measure of uncertainty in the estimation of the reconstructed spectral vector, $\widehat{\mathbf{x}}$ ($\mathbf{x}_r \bigcup \widehat{\mathbf{x}}_u$), is calculated as

$$\widehat{\boldsymbol{\Theta}}_{\widehat{\mathbf{x}}} = \sum_{k=1}^{K} P(k \mid \mathbf{x}_r)\left\{\left(\begin{bmatrix} \mathbf{x}_r \\ \widehat{\boldsymbol{\mu}}_{u,k} \end{bmatrix} - \boldsymbol{\mu}_k\right) \cdot \left(\begin{matrix} \mathbf{x}_r \\ \widehat{\boldsymbol{\mu}}_{u,k} \end{matrix} - \boldsymbol{\mu}_k\right)^T + \begin{bmatrix} 0 & 0 \\ 0 & \widehat{\boldsymbol{\Theta}}_{u,k} \end{bmatrix}\right\} \tag{16.12}$$

where

$$\widehat{\mathbf{\Theta}}_{u,k} = \mathbf{\Theta}_{uu,k} - \mathbf{\Theta}_{ur,k}\mathbf{\Theta}_{rr,k}^{-1}\mathbf{\Theta}_{ru,k}.$$

Equation (16.12) is based on the idea of adapting the trained GMM using the reconstructed spectral vector as an incomplete observation [83]. Even though $\mathbf{y}_r$ is considered reliable during feature reconstruction, the above equation associates a positive, albeit small, measure of uncertainty to it. This helps the uncertainty transformation model to learn the subsequent transformation of these quantities to the cepstral domain, since cepstral uncertainties depend on both $\mathbf{x}_r$ and $\mathbf{x}_u$. If a diagonal covariance matrix is used to model the speech prior, Equation (16.11) and Equation (16.12) can be modified to [66]

$$\widehat{\mathbf{x}}_{u,k} = \sum_{k=1}^{K} P(k \mid \mathbf{x}_r)\boldsymbol{\mu}_{u,k}, \tag{16.13}$$

$$\hat{\boldsymbol{\theta}}_{\widehat{\mathbf{x}}} = \sum_{k=1}^{K} P(k \mid \mathbf{x}_r)\left\{\left(\begin{bmatrix}\mathbf{x}_r \\ \widehat{\mathbf{x}}_{u,k}\end{bmatrix} - \boldsymbol{\mu}_k\right)^2 + \begin{bmatrix}0 \\ \boldsymbol{\theta}_{u,k}\end{bmatrix}\right\}, \tag{16.14}$$

where squaring is done per element of the vector. $\hat{\boldsymbol{\theta}}_{\widehat{\mathbf{x}}}$ and $\boldsymbol{\theta}_{u,k}$ denote the measure of uncertainty in estimation of $\widehat{\mathbf{x}}$ and the unreliable components of the variance of the *k*th Gaussian, respectively. This simplification is due to the fact that all the cross-covariance terms will have the value 0 when the covariance matrix is diagonal. The use of a diagonal covariance matrix reduces the training time and simplifies the calculations.

To perform ASR, the uncertainty transform approach converts the enhanced spectral feature ($\widehat{\mathbf{x}}$) to the cepstral domain. This is straightforward as we have a fully reconstructed feature vector. The main step is to transform the estimated uncertainties to the cepstral domain. In Srinivasan and Wang [71], regression trees are trained to perform this transformation as the true parametric form of this relationship is unknown. If we assume that the cepstral features consist of 39 MFCCs (including the delta and acceleration coefficients), and that the ASR module is based on HMMs that use Gaussians with diagonal covariance matrices to model the observation probability, the goal of the transformation is to estimate the squared difference, $\boldsymbol{\theta}_{\hat{\mathbf{z}}}$, between the reconstructed cepstra, $\hat{\mathbf{z}}$, and the corresponding clean cepstra, $\mathbf{z}$ [15]. The input to the system is the estimated spectral variance ($\hat{\boldsymbol{\theta}}_{\widehat{\mathbf{x}}}$ or $diag(\widehat{\mathbf{\Theta}}_{\widehat{\mathbf{x}}})$, depending on whether diagonal or full covariance matrices are used by the feature reconstruction module). Srinivasan and Wang additionally use the reconstructed cepstral values corresponding to that frame, a preceding frame and a succeeding frame, as input features as they were found to be useful in learning the transformation. The cepstral uncertainties of each of the 39 dimensions are learned using separate regression trees.

Having obtained the enhanced cepstral features and the associated uncertainties, ASR is performed in an uncertainty decoding framework. Since we only have access to the enhanced cepstra, $\hat{\mathbf{z}}$, the observation probability in an HMM-based decoder is calculated by integrating over all possible clean speech cepstral values, $\mathbf{z}$, as shown below:

$$\int_{-\infty}^{\infty} p(\mathbf{z} \mid q, k)p(\hat{\mathbf{z}} \mid \mathbf{z})\mathrm{d}\mathbf{z} = \mathcal{N}(\hat{\mathbf{z}}; \boldsymbol{\mu}_{q,k}, \boldsymbol{\theta}_{q,k} + \boldsymbol{\theta}_{\hat{\mathbf{z}}}). \tag{16.15}$$

In the equation, q denotes a state in the HMM and k indexes the Gaussians used to model the observation probability. $\boldsymbol{\mu}_{q,k}$ and $\boldsymbol{\theta}_{q,k}$ are the corresponding mean and the variance vector

Table 16.1 Word error rates (WER) of the uncertainty transform and the multiple prior based uncertainty transform methods, as well as the reconstruction-based approach. Baseline results of directly recognizing the noisy speech are also shown. MP abbreviates multiple priors. The last column shows the average WER of each of the systems across all the noise types. Reproduced by permission of Narayanan *et al*. [52] © 2011 IEEE.

	Test Set						
System	Car	Babble	Restaurant	Street	Airport	Train	**Average**
Baseline	44.9	43.7	43.2	52.0	44.1	55.2	47.2
Reconstruction	21.5	38.5	42.6	41.5	41.5	39.4	37.5
Uncertainty decoding	18.9	34.2	41.2	40.6	37.0	39.0	35.2
MP reconstruction	19.6	34.8	41.0	38.3	41.1	36.5	35.2
MP uncertainty decoding	18.4	32.8	39.1	37.4	36.9	36.5	33.5

of the *k*th Gaussian. If the observation probability is modeled using Gaussians and if the enhancement is unbiased, this probability can be calculated as shown in the equation [15]. Essentially, the learned variance of a Gaussian component is modified during the recognition stage by adding the estimated cepstral uncertainty to it.

An extension to Srinivasan and Wang's uncertainty transform framework was recently proposed by Narayanan et al. [52]. They propose using multiple prior models of speech, instead of a single large GMM, to better model spectral features. Specifically, they train prior models based on the voicing characteristic of speech by splitting the training data into voiced and unvoiced speech. While reconstructing a noisy spectrogram, frames that are detected as voiced by their voiced/unvoiced (V/UV) detection module are reconstructed using the voiced prior model. Similarly, unvoiced frames are reconstructed using the unvoiced prior model. The V/UV detector is implemented as a binary decision problem, using GMMs to model the underlying density of voiced and unvoiced frames. Like in the uncertainty transform model of Srinivasan and Wang, reconstructed spectral vectors and their corresponding uncertainties are finally transformed to the cepstral domain, and recognition is performed in the uncertainty decoding framework.

The word error rates obtained using the uncertainty transform and the extension by Narayanan *et al*. [52] on the Aurora-4 5000 word closed vocabulary speech recognition task [56] are shown in Table 16.1. This task is based on the *Wall Street Journal* (WSJ0) database. The IBM is estimated using a simple spectral subtraction based approach [71]; the spectral energy in the first and last 50 frames is averaged to create an estimate of the noise spectrum, which is then simultaneously used to 'clean' the noisy spectrogram and to estimate the IBM by comparing it with the energy in each T-F unit. From the table, we can see that, compared to the baseline, uncertainty transform clearly reduces the word error rate in all of the testing conditions. An average improvement of 12 percentage points is obtained over the baseline of directly recognizing noisy speech. Compared to feature reconstruction, an improvement of 2.3 percentage points is obtained. Using multiple prior models further improves the average performance by 1.7 percentage points.

The results show that the uncertainty transform and the use of multiple prior models are effective in dealing with noisy speech utterances. One of the main advantages of the uncertainty transformation is that it enables CASA-based speech enhancement techniques that operate in the spectral domain to be used as a front-end for uncertainty decoding based ASR strategies. The supervised transformation technique can be used whenever the enhancement and the recognition modules operate in different domains. Uncertainty transform techniques provide a clear alternative to missing-data and reconstruction approaches to robust ASR.

16.6 Concluding Remarks

In this chapter, we have discussed facets of CASA and how it can be coupled with ASR to deal with speech recognition in noisy environments. To recapitulate, we discussed perceptual mechanisms that allow humans to analyze the auditory scene. We then looked at how such mechanisms are incorporated in computational models with the goal of achieving human-like performance. Most of the systems discussed in the chapter try to estimate the ideal binary mask, which is an established goal of CASA. Finally, in Section 16.5, we described how CASA can be integrated with ASR.

Although clear advances have been made in the last few years in improving CASA and ASR, challenges remain. CASA challenges lie in developing effective strategies to sequentially organize speech and to deal with unvoiced speech. Apart from additive noise, recent studies have started addressing room reverberation [20,33]. Advances in CASA will have a direct impact on ASR. ASR systems have been demonstrated to perform excellently when the IBM is used. Improvements in IBM estimation will lead to more robust ASR. Over the last decade, attempts at integrating CASA and ASR have yielded fruitful results. Strategies like missing-data ASR, uncertainty transform, and missing feature reconstruction go beyond using CASA as preprocessor for ASR. Further progress in robust ASR can be expected from even tighter coupling between CASA and ASR.

Achieving human-level performance has been the hallmark of many AI endeavors. In CASA, this translates to a meaningful description of the acoustic world. Therefore, recognizing speech in realistic environments is a major benchmark of CASA. Our understanding of how we analyze the auditory scene may eventually pave the way to truly robust ASR.

Acknowledgment

Preparation of this chapter was supported in part by an AFOSR grant (FA9550-08-1-0155).

References

[1] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, “Determination of the potential benefit of time-frequency gain manipulation,” *Ear and Hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[2] P. Assmann and Q. Summerfield, “The perception of speech under adverse acoustic conditions,” in *Speech Processing in the Auditory System*, series on Springer Handbook of Auditory Research, S. Greenberg, W. A. Ainsworth, A. N. Popper, and R. R. Fay, Eds. Berlin: Springer-Verlag, 2004, vol. 18.

[3] J. Barker, L. Josifovski, M. P. Cooke, and P. D. Green, “Soft decisions in missing data techniques for robust automatic speech recognition.” in *Proceedings of the International Conference on Spoken Language Processing*, Beijing, China, 2000, pp. 373–376.

[4] J. Barker, M. P. Cooke, and D. P. W. Ellis, “Decoding speech in the presence of other sources,” *Speech Communication*, vol. 45, pp. 5–25, 2005.

[5] M. Berouti, R. Schwartz, and R. Makhoul, “Enhancement of speech corrupted by acoustic noise,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1979.

[6] P. Boersma and D. Weenink. (2002) Praat: Doing phonetics by computer, version 4.0.26. Available at: http://www.fon.hum.uva.nl/praat.

[7] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, pp. 113–120, 1979.

[8] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[9] G. J. Brown and M. P. Cooke, “Computational auditory scene analysis,” *Computer Speech & Language*, vol. 8, pp. 297–336, 1994.

[10] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, “Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask,” *Journal of Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[11] E. C. Cherry, “Some experiments on recognition of speech, with one and with two ears,” *Journal of Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[12] E. C. Cherry, *On Human Communication*. Cambridge, MA: MIT Press, 1957.

[13] M. P. Cooke, P. Greene, L. Josifovski, and A. Vizinho, “Robust automatic speech recognition with missing and uncertain acoustic data,” *Speech Communication*, vol. 34, pp. 141–177, 2001.

[14] A. de Cheveigne, “Multiple F0 estimation.” in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley-IEEE Press, 2006, pp. 45–80.

[15] L. Deng, J. Droppo, and A. Acero, “Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 412–421, 2005.

[16] N. I. Durlach, “Note on the equalization and cancellation theory of binaural masking level differences,” *Journal of Acoustical Society of America*, vol. 32, no. 8, pp. 1075–1076, 1960.

[17] M. El-Maliki and A. Drygajlo, “Missing features detection and handling for robust speaker verification,” in *Proceedings of Interspeech*, 1999, pp. 975–978.

[18] B. R. Glasberg and B. C. J. Moore, “Derivation of auditory filter shapes from notched-noise data,” *Hearing Research*, vol. 47, pp. 103–138, 1990.

[19] K. Han and D. L. Wang, “An SVM based classification approach to speech separation.” in *Proccedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4632–4635.

[20] S. Harding, J. Barker, and G. J. Brown, “Mask estimation for missing data speech recognition based on statistics of binaural interaction,” *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 58–67, 2006.

[21] W. Hartmann and E. Fosler-Lussier, “Investigations into the incorporation of the ideal binary mask in ASR,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4804–4807.

[22] H. Helmholtz, *On the Sensation of Tone*, 2nd ed. New York: Dover Publishers, 1863.

[23] H. G. Hirsch and C. Ehrlicher, “Noise estimation techniques for robust speech recognition,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1995, pp. 153–156.

[24] G. Hu and D. L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[25] G. Hu and D. L. Wang, “Auditory segmentation based on onset and offset analysis,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 396–405, 2007.

[26] G. Hu and D. L. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, pp. 2067–2079, 2010.

[27] G. Hu and D. L. Wang, “Speech segregation based on pitch tracking and amplitude modulation.” in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 79–82.

[28] G. Hu and D. L. Wang, “Segregation of unvoiced speech from nonspeech interference,” *Journal of Acoustical Society of America*, vol. 124, pp. 1306–1319, 2008.

[29] K. Hu and D. L. Wang, “Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction.” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 6, pp. 1600–1609, 2011.

[30] K. Hu and D. L. Wang, “Unsupervised sequential organization for cochannel speech separation.” in *Proceedings of Interspeech*, Makuhari, Japan, 2010, pp. 2790–2793.

[31] L. A. Jeffress, “A place theory of sound localization,” *Comparative Physiology and Psychology*, vol. 41, pp. 35–39, 1948.

[32] Z. Jin and D. L. Wang, “A supervised learning approach to monaural segregation of reverberant speech,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 625–638, 2009.

[33] Z. Jin and D. L. Wang, “HMM-based multipitch tracking for noisy and reverberant speech,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1091–1102, 2011.

[34] L. Josifovski, M. Cooke, P. Green, and A. Vizihno, “State based imputation of missing data for robust speech recognition and speech enhancement,” in *Proceedings of Interspeech*, 1999, p. 2837–2840.

[35] S. G. Karadogan, J. Larsen, M. S. Pedersen, and J. B. Boldt, “Robust isolated speech recognition using binary masks.” in *Proceedings of the European Signal Processing Conference*, 2010, pp. 1988–1992.

[36] G. Kim, Y. Lu, Y. Hu, and P. Loizou, “An algorithm that improves speech intelligibility in noise for normal-hearing listeners,” *Journal of Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[37] W. Kim and R. Stern, “Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise,” *Speech Communication*, vol. 53, pp. 1–11, 2011.

[38] A. Klapuri, “Multipitch analysis of polyphonic music and speech signals using an auditory model,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 16, no. 2, pp. 255–266, 2008.

[39] B. Kollmeier and R. Koch, “Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction,” *Journal of Acoustical Society of America*, vol. 95, pp. 1593–1602, 1994.

[40] A. Korthauer, “Robust estimation of the SNR of noisy speech signals for the quality evaluation of speech databases,” in *Proceedings of ROBUST’99 Workshop*, 1999, pp. 123–126.

[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1998.

[42] N. Li and P. C. Loizou, “Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction,” *Journal of Acoustical Society of America*, vol. 123, no. 3, pp. 1673–1682, 2008.

[43] Y. Li and D. L. Wang, “On the optimality of ideal binary time-frequency masks,” *Speech Communication*, vol. 51, pp. 230–239, 2009.

[44] J. C. R. Licklider, “A duplex theory of pitch perception,” *Experimentia*, vol. 7, pp. 128–134, 1951.

[45] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, “The precedence effect,” *Journal of Acoustical Society of America*, vol. 106, pp. 1633–1654, 1999.

[46] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, Florida: CRC Press, 2007.

[47] N. Ma, P. Green, J. Barker, and A. Coy, “Exploiting correlogram structure for robust speech recognition with multiple speech sources,” *Speech Communication*, vol. 49, pp. 874–891, 2007.

[48] R. Meddis, M. J. Hewitt, and T. M. Shackelton, “Implementation details of a computational model of the inner hair-cell/auditory-nerve synapse,” *Journal of Acoustical Society of America*, vol. 122, no. 2, pp. 1165–1172, 1990.

[49] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th ed. London, UK: Academic Press, 2003.

[50] A. Narayanan and D. L. Wang, “Robust speech recognition from binary masks,” *Journal of Acoustical Society of America*, vol. 128, pp. EL217–222, 2010.

[51] A. Narayanan and D. L. Wang, “On the use of ideal binary masks to improve phone classification.” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processsing*, 2011, pp. 5212–5215.

[52] A. Narayanan, X. Zhao, D. L. Wang, and E. Fosler-Lussier, “Robust speech recognition using multiple prior models for speech reconstruction,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 4800–4803.

[53] E. Nemer, R. Goubran, and S. Mahmoud, “SNR estimation of speech signals using subbands and fourth-order statistics,” *IEEE Signal Processing Letters*, vol. 6, no. 7, pp. 504–512, 1999.

[54] S. Nooteboom, “The prosody of speech: Melody and rhythm,” in *The Handbook of Phonetic Science*, W. J. Hardcastle and J. Laver, Eds. Blackwell: Oxford, UK, 1997, pp. 640–673.

[55] K. J. Palomaki, G. J. Brown, and D. L. Wang, “A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation,” *Speech Communication*, vol. 43, pp. 361–378, 2004.

[56] N. Parihar and J. Picone, “Analysis of the Aurora large vocabulary evalutions,” in *Proceedings of the European Conference on Speech Communication and Technology*, 2003, pp. 337–340.

[57] T. W. Parsons, “Separation of speech from interfering speech by means of harmonic selection,” *Journal of Acoustical Society of America*, vol. 60, no. 4, pp. 911–918, 1976.

[58] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammatone function,” Technical Report 2341, MRC Applied Psychology Unit, Cambridge, UK, 1988.

[59] B. Raj and R. Stern, “Missing-feature approaches in speech recognition,” *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

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

[61] P. Renevey and A. Drygajlo, “Detection of reliable features for speech recognition in noisy conditions using a statistical criterion,” in *Proceedings of Consistent & Reliable Acoustic Cues for Sound Analysis Workshop*, 2001, pp. 71–74.

[62] N. Roman, D. L. Wang, and G. J. Brown, “Speech segregation based on sound localization,” *Journal of Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[63] S. T. Roweis, “One microphone source separation,” in *Advances in Neural Information Processing System 13*, 2000, pp. 793–799.

[64] M. R. Schroeder, “Period histogram and product spectrum: New methods for fundamental-frequency measurement,” *Journal of Acoustical Society of America*, vol. 43, pp. 829–834, 1968.

[65] M. Seltzer, B. Raj, and R. Stern, “A bayesian classifer for spectrographic mask estimation for missing feature speech recognition,” *Speech Communication*, vol. 43, no. 4, pp. 379–393, 2004.

[66] Y. Shao, “Sequential organization in computational auditory scene analysis,” PhD dissertation, The Ohio State Univeristy, 2007.

[67] Y. Shao and D. L. Wang, “Model-based sequential organization in cochannel speech,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 289–298, 2006.

[68] P. Y. Simard, D. Steinkraus, and J. C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in *Proceedings of the International Conference on Document Analysis and Recognition*, 2003, pp. 958–963.

[69] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[70] W. Speith, J. F. Curtis, and J. C. Webseter, “Responding to one of two simultaneous messages,” *Journal of Acoustical Society of America*, vol. 26, pp. 391–396, 1954.

[71] S. Srinivasan and D. L. Wang, “Transforming binary uncertainties for robust speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 2130–2140, 2007.

[72] S. Srinivasan and D. L. Wang, “Robust speech recognition by integrating speech separation and hypothesis testing,” *Speech Communication*, vol. 52, pp. 72–81, 2010.

[73] S. Srinivasan, N. Roman, and D. L. Wang, “Binary and ratio time-frequency masks for robust speech recognition,” *Speech Communication*, vol. 48, pp. 1486–1501, 2006.

[74] J. Tchorz and B. Kollmeier, “SNR estimation based on amplitude modulation analysis with applications to noise suppression,” *IEEE Transactions on Audio, Speech and Signal Processing*, vol. 11, pp. 184–192, 2003.

[75] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, “Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study,” in *Proceedings of Interspeech*, 1999, pp. 2407–2410.

[76] D. L. Wang, “On ideal binary masks as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Kluwer Academic, 2005, pp. 181–197.

[77] D. L. Wang, “Time-frequency masking for speech separation and its potential for hearing aid design,” *Trends in Amplification*, vol. 12, pp. 332–353, 2008.

[78] D. L. Wang and G. J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[79] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.

[80] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, “Speech perception of noise with binary gains,” *Journal of Acoustical Society of America*, vol. 124, no. 4, pp. 2303–2307, 2008.

[81] D. L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, “Speech intelligibility in background noise with ideal binary time-frequency masking,” *Journal of Acoustical Society of America*, vol. 125, pp. 2336–2347, 2009.

[82] M. Weintraub, “A theory and computational model of auditory monaural sound separation,” PhD dissertation, Stanford University, 1985.

[83] D. Williams, X. Liao, Y. Xue, and L. Carin, “Incomplete-data classification using logistic regression,” in *Proceedings of the 22nd Internation Conference on Machine Learning*, 2005, pp. 972–979.

[84] J. Woodruff and D. L. Wang, "Sequential organization of speech in reverberant environments by integrating monaural grouping and binaural localization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, pp. 1856–1866, 2010.

[85] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[86] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.