12

An Auditory Scene Analysis Approach to Monaural Speech Segregation

Guoning Hu[1] and DeLiang Wang[2]

[1] The Ohio State University, Biophysics Program
Columbus, OH 43210

[2] The Ohio State University, Department of Computer Science & Engineering and
Center for Cognitive Science
Columbus, OH 43210

A human listener has the remarkable ability to segregate an acoustic mixture and attend to a target sound. This perceptual process is called auditory scene analysis (ASA). Moreover, the listener can accomplish much of auditory scene analysis with only one ear. Research in ASA has inspired many studies in computational auditory scene analysis (CASA) for sound segregation. In this chapter we introduce a CASA approach to monaural speech segregation. After a brief overview of CASA, we present in detail a CASA system that segregates both voiced and unvoiced speech. Our description covers the major stages of CASA, including feature extraction, auditory segmentation, and grouping.

12.1 Introduction

We live in an environment rich in sound from many sources. The presence of multiple sound sources complicates the processing of the target sound we are interested in, and often causes serious problems for many applications, such as automatic speech recognition and voice communication. There has been extensive effort to develop computational systems that automatically separate target sound or attenuate background interference. When target and interference come from different directions and multiple microphones are available, one may remove interference using spatial filtering that extracts the signal from the target direction or cancels the signals from the interfering directions [29], or independent component analysis [26]. These approaches do not apply to the situations when target and interference originate from the same direction or only mono-recordings are available. In the monaural (one microphone)

Reprinted from *Topics in Acoustic Echo and Noise Control*, E. Hänsler and G. Schmidt (Eds.), Springer, New York, 2006, pp. 485-515.

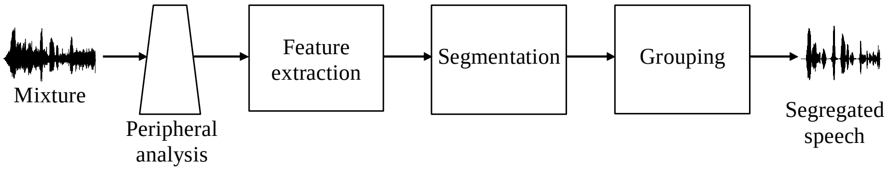

Fig. 12.1. Schematic diagram of a typical CASA system.

situation, one must consider the intrinsic properties of target or interference to distinguish and separate them.

As a special case of monaural separation, monaural speech segregation is of particular importance. Here a major challenge is the variety of interference; the interference can change in time and space in an unpredictable manner. For decades, various methods have been proposed for monaural speech enhancement, such as spectral subtraction [5], subspace analysis [17], hidden Markov modeling [46], and sinusoidal modeling [28]. These methods usually assume certain properties (or models) of interference and then enhance speech or attenuate interference based on these assumptions. Their capacity for dealing with the variability of interference is much limited in comparison with human speech segregation. This contrast has motivated a different approach to monaural speech segregation – mimicking the auditory process of source separation.

The auditory segregation process is termed by Bregman as *auditory scene analysis (ASA)* [6], which is considered to take place in two main stages: Segmentation and grouping. In segmentation, the acoustic input is decomposed into segments or sensory elements, each of which should originate from a single source. In grouping, the segments that likely arise from the same source are grouped together. Segmentation and grouping are guided by perceptual principles that determine how the auditory scene is organized according to ASA cues. These cues characterize intrinsic sound properties, including harmonicity, onset and offset, location, and prior knowledge of specific sounds.

Research in ASA has inspired considerable work to build CASA (computational auditory scene analysis) systems for sound segregation (for reviews see [44, 8]). A main advantage is that CASA does not make strong assumptions about interference. A typical CASA system is shown in Fig. 12.1. It contains four stages: Peripheral analysis, feature extraction, segmentation, and grouping. The peripheral processing decomposes the auditory scene into a time-frequency (T-F) representation via bandpass filtering and time windowing. The second stage extracts auditory features corresponding to ASA cues, which will be used in subsequent segmentation and grouping. In segmentation and grouping, the system generates segments for both target and interference and groups the segments originating from the target into a target stream. A stream corresponds to a sound source. The waveform of segregated target can then be resynthesized from the target stream [53, 7, 52].

As an illustration, Figs. 12.2(a) and 12.2(b) show a T-F decomposition and the waveform of a male utterance, "Her right hand aches whenever the barometric pressure changes," from the TIMIT database [18]. Figs. 12.2(c) and 12.2(d) show a T-F decomposition and the waveform of this utterance mixed with a crowd noise in playground, at the overall SNR of 0 dB. Here the input is decomposed using a filterbank with 128 gammatone filters [36] and 20-ms rectangular time windows with 10-ms window shift (see Sec. 12.3 for implementation details). The small T-F area within each filter channel and time window is referred to as a T-F unit. Figs. 12.2(a) and 12.2(c) show the energy within each T-F unit, where a brighter pixel indicates stronger energy. Fig. 12.2(e) shows the target stream we aim to segregate, which contains all the T-F units dominated by the target. To obtain this stream, a typical CASA system first merges neighboring T-F units dominated by target speech into segments, shown as the contiguous black regions in the figure, in the stage of segmentation. In this stage, the system may also generate segments for interference. Then in the stage of grouping, the system determines for each segment whether it belongs to the target and groups them accordingly. Fig. 12.2(f) shows the waveform resynthesized from the target stream in Fig. 12.2(e).

Brown and Wang have recently written a review chapter on CASA for speech segregation, also included in a Springer volume [8]. Instead of another review, this chapter mainly describes our systematic effort on monaural speech segregation. The chapter is organized as follows. In Sec. 12.2, we give a brief overview of other CASA studies on monaural speech segregation. We then describe in depth the major stages of our CASA system in the subsequent four sections. Sec. 12.7 concludes the chapter.

12.2 Computational Auditory Scene Analysis

Natural speech contains both voiced and unvoiced portions. Voiced speech is periodic or quasi-periodic. Periodicity and temporal continuity are two major ASA cues for voiced speech. A well-established representation for periodicity and pitch perception is a correlogram - a running autocorrelation of each filter response across an auditory filterbank [31, 48]. The correlogram has been adopted by many CASA systems for monaural segregation of voiced speech [53, 13, 7, 16, 52, 23]. In what is regarded as the first CASA model, Weintraub used a coincidence function, a version of autocorrelation, to capture periodicity as well as amplitude modulation (AM) [53]. He then used the coincidence function to track pitch contours of multiple utterances. Sounds from different speakers are separated by using iterative spectral estimation according to pitch and temporal continuity. Cooke proposed a model that first generates local segments based on filter response frequencies and temporal continuity [13]. These segments are merged into groups based on common harmonicity and common AM. A pitch contour is then obtained for each group, and groups

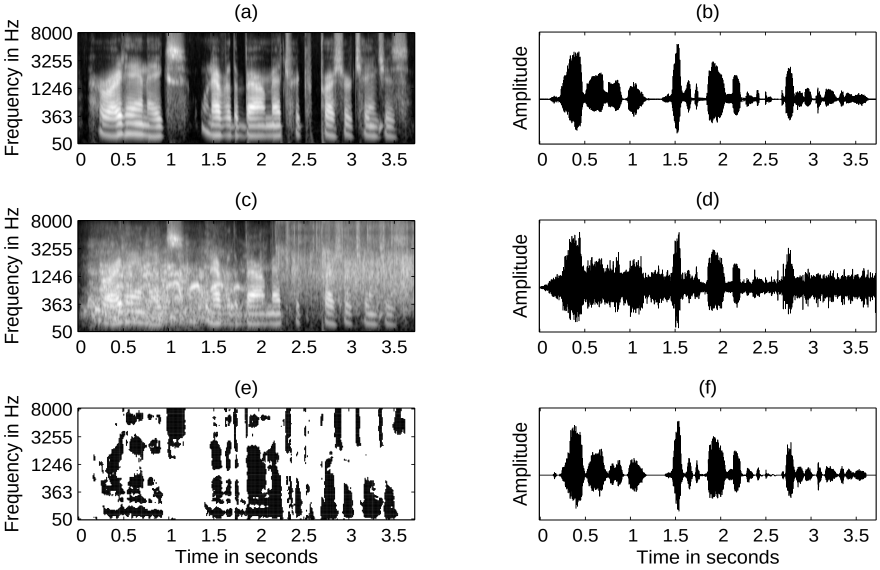

Fig. 12.2. Signal representation. (a) T-F decomposion of a male utterance, "Her right hand aches whenever the barometric pressure changes." (b) Waveform of the utterance. (c) T-F decomposition of the utterance mixed with a crowd noise in playground. (d) Waveform of the mixture. (e) Target stream composed of all the T-F units (black regions) dominated by the target (ideal binary mask). (f) The waveform resynthesized from the target stream.

with similar pitch contours are put into the same stream. Brown and Cooke proposed to form segments based on correlation of filter responses across frequency and frequency transition across time [7]. These segments are grouped by common periodicity and common onset and offset. Wang and Brown used a two-layer oscillator network for speech segregation [52]. In the first layer, segments are formed based on cross-channel correlation and temporal continuity. In the second layer, segments are grouped into two streams, one for the target and the other its background on the basis of dominant pitch in each time frame. The above systems are mainly data-driven approaches. Ellis developed a prediction-driven system which generates predictions using a world model and compares the predictions against the input [16]. The world model includes three types of sound elements: Noise cloud, transient click, and harmonic sound.

12.2.1 Computational Goal of CASA

A critical issue in developing a CASA system is to determine its computational goal [32]. With the initial analysis into T-F units described in Sec. 12.1, we have suggested that the computational goal of CASA should be to retain the

T-F units where target speech is more intense than interference and remove others [21, 23]. In other words, the goal is to identify a binary T-F mask, referred to as the *ideal binary mask*, where 1 indicates that target is stronger than interference in the corresponding T-F unit and 0 otherwise. Target speech can then be resynthesized with the ideal mask by retaining the acoustic energy from T-F regions corresponding to 1's and rejecting other energy. This computational goal is supported by the auditory masking phenomenon: Within a critical band, a weaker signal tends to be masked by a stronger one [35]. In addition, there is considerable evidence supporting the ideal binary mask as the CASA objective from both human speech intelligibility [42, 12, 9] and automatic speech recognition [14, 42] studies (for an extensive discussion see [51]). What Fig. 12.2(e) shows, in fact, is an ideal binary mask for the mixture in Fig. 12.2(c). As shown in Fig. 12.2(f), the speech resynthesized from the ideal binary mask is close to the original clean utterance in Fig. 12.2(b).

12.2.2 Motivation and Overview of the Approach

A common problem in earlier CASA systems is that they do not separate voiced speech well in the high-frequency range from broadband interference. This problem is closely related to the peripheral analysis of the input scene. Most CASA systems perform initial frequency analysis with an auditory filterbank, where the bandwidth of a filter increases quasi-logarithmically with its center frequency. These filters are usually derived from psychophysical data and mimic cochlear filtering. An important observation is that the structure of cochlear filtering limits the ability of human listeners to resolve harmonics [38, 40]. In the low-frequency range, harmonics are resolved since the corresponding auditory filters have narrow passbands including only one harmonic. In the high-frequency range, harmonics are generally unresolved since the corresponding auditory filters have wide passbands including multiple harmonics. In addition, psychophysical evidence suggests that the human auditory system processes resolved and unresolved harmonics differently [11, 3]. Hence, one should carefully consider the distinctions between resolved and unresolved harmonics. The earlier CASA systems employ the same strategy to segregate all the harmonics, which works reasonably well for resolved harmonics but poorly for unresolved ones.

A basic fact of acoustic interaction is that the filter responses to multiple harmonics are amplitude-modulated and the response envelopes fluctuate at the fundamental frequency (f_0) of target speech [19]. Fig. 12.3 shows the response and its envelope of a gammatone filter centered at 2.5 kHz within a time frame (from 0.7 s to 0.72 s). The input is the clean utterance in Fig. 12.2(b). The response in Fig. 12.3 is strongly amplitude-modulated, and its envelope fluctuates at the f_0 rate in this frame.

Motivated by the above considerations, we have proposed to employ different methods to segregate resolved and unresolved harmonics of target speech

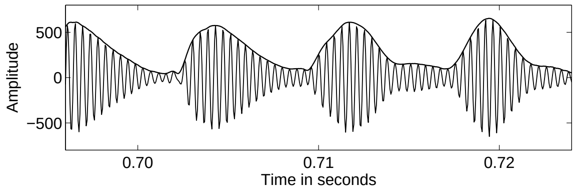

Fig. 12.3. AM effects for filter responses to multiple harmonics. The input is the utterance in Fig. 12.2(b). The filter is centered at 2.5 kHz.

[23]. For resolved harmonics, we generate segments based on temporal continuity and cross-channel correlation, and these segments are grouped according to common periodicity, similar to [52]. For unresolved harmonics, we generate segments based on common AM in addition to temporal continuity. These segments are further grouped based on AM rates, which are obtained from the temporal fluctuations of the corresponding response envelopes.

So far the discussion is focused on voiced speech. Compared with voiced speech, unvoiced speech is generally much weaker and more susceptible to interfering sounds. In addition, unvoiced speech lacks harmonic structure and is noise-like itself. As a result, segregating unvoiced speech is significantly more challenging and little previous work has addressed this problem.

We have proposed to segment unvoiced speech based on onset and offset analysis [24]. Onsets and offsets are important ASA cues [6] because different sound sources in an environment seldom start and end simultaneously. In addition, there is strong evidence for onset detection by auditory neurons [37]. In the time domain, onsets and offsets likely form boundaries between sounds from different sources. Common onsets and offsets also provide natural cues to integrate sounds from the same source across frequency. In addition, onset/offset based segmentation is applicable to both voiced and unvoiced speech.

Given segments, the next task is to group segments of unvoiced speech. When interference is non-speech, we may formulate this as a classification task, i.e., to classify segments as unvoiced speech or interference. Since each segment should belong to one source, segments dominated by unvoiced speech are likely to have similar acoustic-phonetic characteristics as those of clean speech, whereas segments dominated by interference are likely to be different. Therefore, we can group segments for unvoiced speech by analyzing their acoustic-phonetic features [25].

In the following sections, we describe our systematic investigation into segregation of both voiced and unvoiced speech. Our model includes all the major stages of a typical CASA system shown in Fig. 12.1.

12.3 Peripheral Analysis and Feature Extraction

We describe below early auditory processing that first decomposes the input in the T-F domain, and then extracts auditory features corresponding to ASA cues.

12.3.1 Auditory Periphery

Cochlear filtering is commonly modeled by a gammatone filterbank that decomposes the input in the frequency domain [36]. The impulse response of a gammatone filter centered at frequency f is:

$$g(f,t) = \begin{cases} b^a t^{a-1} e^{-2\pi bt} \cos(2\pi ft), & t \geq 0, \\ 0, & \text{else}, \end{cases} \tag{12.1}$$

where $a = 4$ is the order of the filter. b is the equivalent rectangular bandwidth, which increases as the center frequency f increases. For a filter channel c, let f_c be the center frequency. Let $x(t)$ be the input signal, the response from channel c, $x(c,t)$, is then

$$x(c,t) = x(t) * g(f_c, t), \tag{12.2}$$

where "$*$" denotes convolution. The response is shifted backwards by $(a-1)/(2\pi b)$ to compensate for the filter delay [20]. We find that this delay compensation gives a small but consistent performance improvement. In addition, the gain of each filter is adjusted according to equal loudness contours [27] in order to simulate the pressure gains of the outer and middle ears.

The response of a filter channel can be further processed by the Meddis model of auditory nerve transduction [33]. This model simulates the nonlinear processes of the auditory nerve, such as rectification, saturation, and phase locking. Its output represents the firing rate of an auditory nerve fiber, denoted by $h(c,t)$.

In each filter channel, the output is divided into 20-ms time frames with 10-ms overlapping between consecutive frames. This frame size is commonly used for speech analysis. Examples of this T-F decomposition are shown in Figs. 12.2(a) and 12.2(c). The resulting time-frequency representation is called a *cochleagram.*

12.3.2 Correlogram and Cross-Channel Correlation

As discussed in Sec. 12.2, a correlogram is a commonly used periodicity representation, which consists of autocorrelations of filter responses across all the filter channels. Let u_{cm} denote a T-F unit for frequency channel c and time frame m, the corresponding normalized autocorrelation of the filter response is given by

$$A_{\mathrm{H}}(c,m,\tau) = \frac{\sum_n h\big(c, mT_{\mathrm{f}} - nT_{\mathrm{s}}\big)\, h\big(c, mT_{\mathrm{f}} - nT_{\mathrm{s}} - \tau T_{\mathrm{s}}\big)}{\sum_n h^2\big(c, mT_{\mathrm{f}} - nT_{\mathrm{s}}\big)}\,. \tag{12.3}$$

Here, τ is the delay and n denotes digitized time. $T_{\mathrm{f}} = 10$ ms, the time shift from one frame to the next and T_{s} is denoting the sampling time. The above summation is over the period of a time frame.

As shown in [7, 52], cross-channel correlation measures the similarity between the responses of two adjacent filter channels and indicates whether the filters respond to the same sound component. For T-F unit u_{cm}, its cross-channel correlation with $u_{c+1,m}$ is given by

$$C_{\mathrm{H}}(c,m) = \sum_{\tau=0}^{L} \widetilde{A}_{\mathrm{H}}(c,m,\tau)\, \widetilde{A}_{\mathrm{H}}(c+1,m,\tau)\,, \tag{12.4}$$

where $\widetilde{A}_{\mathrm{H}}(c,m,\tau)$ denotes $A_{\mathrm{H}}(c,m,\tau)$ normalized to 0 mean and unity variance and $LT_{\mathrm{s}} = 12.5$ ms - the maximum delay for A_{H}.

The AM information is carried by the response envelope. A general way to obtain response envelope is to perform half-wave rectification followed by lowpass filtering. Since we are interested in the envelope fluctuations corresponding to target pitch, here we perform a bandpass filtering instead, where the passband corresponds to the plausible f_0 range of target speech. Let $h_{\mathrm{E}}(c,t)$ denote the resulting envelope.

Similar to Eqs. 12.3 and 12.4, we can compute a normalized envelope autocorrelation to represent AM rates:

$$A_{\mathrm{E}}(c,m,\tau) = \frac{\sum_n h_{\mathrm{E}}\big(c, mT_{\mathrm{f}} - nT_{\mathrm{s}}\big)\, h_{\mathrm{E}}\big(c, mT_{\mathrm{f}} - nT_{\mathrm{s}} - \tau T_{\mathrm{s}}\big)}{\sum_n h_{\mathrm{E}}^2\big(c, mT_{\mathrm{f}} - nT_{\mathrm{s}}\big)} \tag{12.5}$$

and cross-channel correlation of response envelopes,

$$C_{\mathrm{E}}(c,m) = \sum_{\tau=0}^{L} \widetilde{A}_{\mathrm{E}}(c,m,\tau)\, \widetilde{A}_{\mathrm{E}}(c+1,m,\tau)\,. \tag{12.6}$$

Figs. 12.4(a) and 12.4(b) illustrate the correlogram and the envelope correlogram as well as the cross-channel correlation at time frame 70 (i.e., 0.7 s from the beginning of the stimulus) for the utterance in Fig. 12.2(b), and Figs. 12.4(c) and 12.4(d) the corresponding responses to the mixture in Fig. 12.2(d). As shown in the figure, the autocorrelation of filter response generally reflects the periodicity of a single harmonic for a channel in the low-frequency range where harmonics are resolved. The autocorrelation is amplitude-modulated in high-frequency channels where harmonics are unresolved. As a result, these autocorrelations are not as highly correlated between adjacent channels. On the other hand, the corresponding autocorrelations of response envelopes are more correlated, as shown in the cross-channel correlations of response envelopes, since they have similar fluctuation patterns.

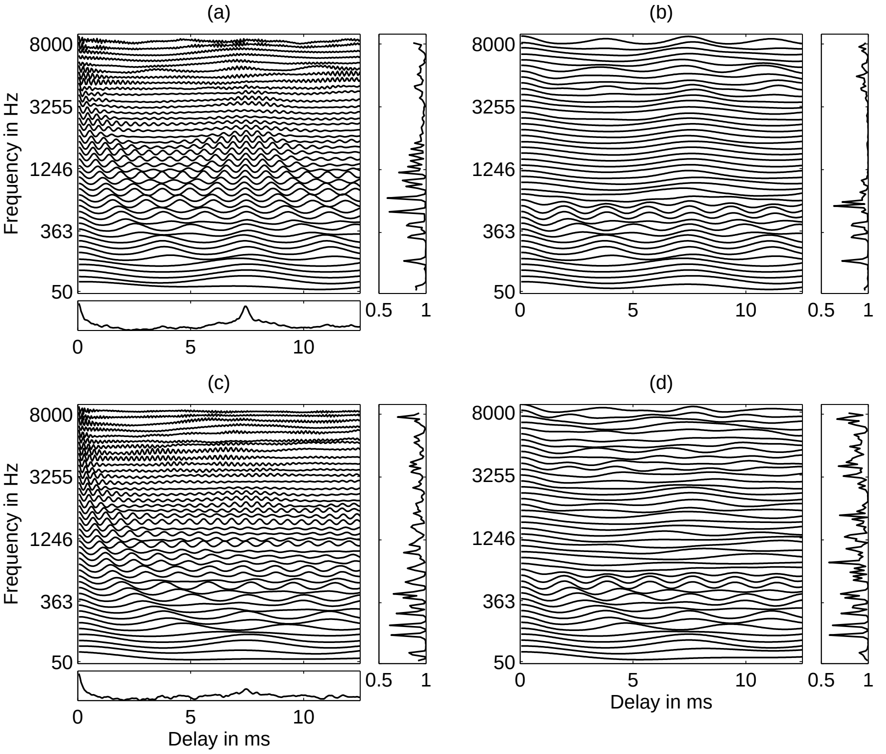

Fig. 12.4. Auditory features. (a) Correlogram at frame 70 (i.e. 0.7 second after the onset) for the utterance in Fig. 12.2(b). For clarity, every third channel is displayed. The corresponding cross-channel correlation is given in the right panel, and the summary correlogram in the bottom panel. (b) Envelope correlogram for the utterance. The corresponding cross-channel envelope correlation is shown in the right panel. (c) Correlogram and cross-channel correlation for the mixture in Fig. 12.2(d). (d) Envelope correlogram and cross-channel envelope correlation for the mixture.

12.3.3 Onset and Offset

Onsets and offsets correspond to sudden amplitude increases and decreases. A standard way to identify such intensity changes is to take the first-order derivative of intensity with respect to time and then find the peaks and valleys of the derivative. Because of intrinsic intensity fluctuations, many peaks and valleys of the derivative do not correspond to actual onsets and offsets. To reduce such fluctuations, we smooth the intensity over time, as is commonly done in edge detection for image analysis. The intensity is basically the square of the envelope of filter response. Smoothing can be performed through either a diffusion process [43] or lowpass filtering. Here we consider a special case of Gaussian smoothing. First we calculate the response envelope with half-wave rectification and lowpass filtering. Since here we are interested in low-rate

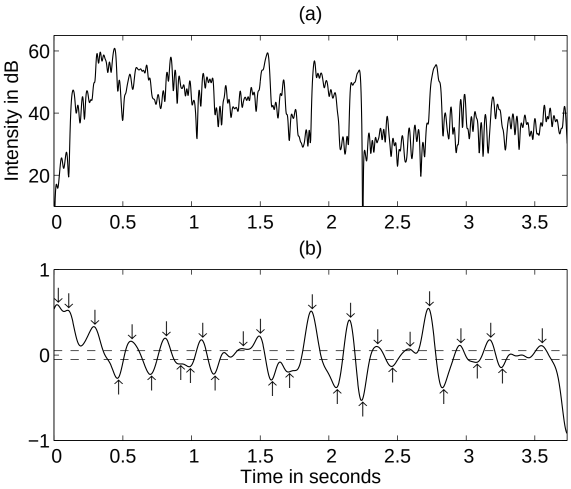

Fig. 12.5. Onset and offset detection. The input is the response of a gammatone filter to the mixture in Fig. 12.2(d). The upper panel shows the response intensity, and the lower panel shows the results of onset and offset detection using Gaussian smoothing ($\sigma = 16$). The threshold for onset detection is 0.05 and for offset detection is -0.05, indicated by the dash lines. Detected onsets are marked by downward arrows and offsets by upward arrows.

fluctuations of envelope, the cutoff frequency of the lowpass filter should be set smaller than 30 Hz. The obtained low-rate envelope is denoted by $x_{\mathrm{E}}(c,t)$. The smoothed intensity is obtained by the convolution of the intensity (in decibels) and a Gaussian kernel with mean 0 and variance σ^2. The derivative of the smoothed response is:

$$\begin{aligned}&\frac{d}{dt}\left\{10\left[\log_{10} x_{\mathrm{E}}^2(c,t)\right] * \left[\frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(-\frac{t^2}{2\sigma^2}\right)\right]\right\}\\ &= -20\log_{10}\left|x_{\mathrm{E}}(c,t)\right| * \left[\frac{t}{\sqrt{2\pi}\,\sigma^3}\exp\left(-\frac{t^2}{2\sigma^2}\right)\right].\end{aligned}$$

Onsets correspond to the peaks of the derivative above a certain threshold, and offsets the valleys below a certain threshold. The purpose of thresholding is to remove peaks and valleys corresponding to insignificant intensity fluctuations. The above procedure is very similar to the standard Canny edge detector in image processing [10]. An example of the above procedure is shown in Fig. 12.5.

12.3.4 Pitch Determination

A periodic sound consists of a harmonic series, each harmonic having a frequency that equals or is a multiple of f_0. A frequently-used method for pitch determination is to simply pool autocorrelations across all the channels and then identify a global peak in the summary correlogram [34]. When a harmonic sound is presented, the autocorrelations of the activated filters in a correlogram all exhibit a peak at the delay corresponding to the pitch period. Let $A_{\mathrm{H}}(m,\tau)$ be the summary correlogram at frame m, that is,

$$A_{\mathrm{H}}(m,\tau) = \sum_{c} A_{\mathrm{H}}(c,m,\tau)\,. \tag{12.7}$$

The estimated pitch period at frame m, $\tau_{\mathrm{S}}(m)$, is the lag corresponding to the maximum of $A_{\mathrm{H}}(m,\tau)$ in the plausible pitch range of target speech. The bottom panels of Figs. 12.4(a) and 12.4(c) shows examples of summary correlogram. The peak at 7.21 ms in Fig. 12.4(c), representing the estimated pitch period, turns out to equal that of target speech (indicated by the peak in Fig. 12.4(a)).

There are several problems with the above method. First, it gives a pitch value at each frame no matter whether the signal at a particular frame is periodic or not. Second, detected pitches in neighboring frames may correspond to different sound sources. Third, it may not give a reliable estimate of target pitch even if it exists, when the signal-to-noise ratio (SNR) is low. This is because the autocorrelations in many channels exhibit peaks not corresponding to the periodicity of the target. To address these problems, we apply the Wang and Brown algorithm [52] in an initial grouping stage. The grouping in their algorithm is based on the dominant pitch of each time frame, and can eliminate many T-F units that unlikely belong to the target. With this initial grouping, we track a *target pitch contour* by pooling autocorrelations from the remaining T-F units. The initial grouping is not accurate in the high-frequency range; however, this stage is employed only for the purpose of pitch tracking. Note that pitch detection requires only a portion of harmonics; the fact that the Wang and Brown algorithm works reasonably well in the low-frequency range accords well with the perceptual evidence that human pitch detection primarily relies on lower harmonics [39]. To deal with the third problem, we take advantage of the pitch continuity to enhance the reliability of target pitch tracking [23]. Specifically, we first determine the reliability of an estimated pitch based on its coherence with the periodicity patterns of the retained T-F units in initial grouping, and then use pitch continuity to interpolate for unreliable pitch points on the basis of reliable ones.

The algorithm given in [23] assumes that the target has a continuous pitch contour throughout the whole utterance. We note that it can be applied iteratively to handle the general situation when the target utterance contains multiple pitch contours separated by unvoiced speech or silence. This is because the initial grouping by the Wang-Brown algorithm is based on the longest segment. Specifically, after extracting the first pitch contour based on the longest segment, the algorithm can then be applied to extract the next longest pitch contour from remaining time frames where no target pitch has been detected. This process can repeat until no more significant pitch contour is detected.

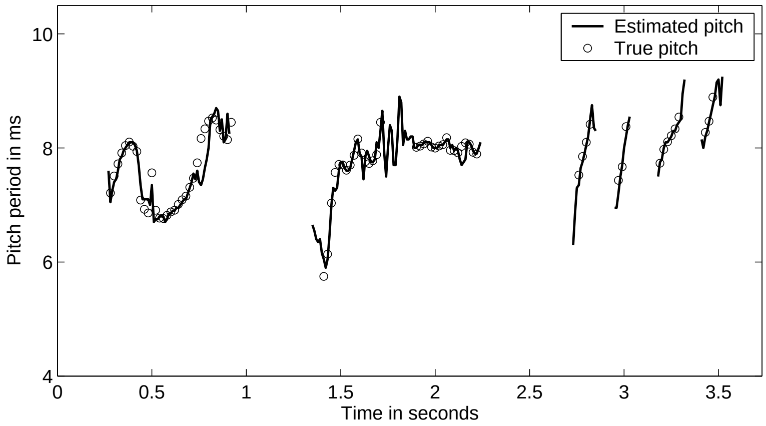

Fig. 12.6. Results of pitch tracking for the mixture in Fig. 12.2(d). Solid lines indicate estimated target pitch contours. True pitch points are marked by circles. For clarity, every other frame is displayed.

However, when interference also contains periodic signal, the above procedure may generate pitch contours for interference as well. To determine the source for each pitch contour is the task of sequential grouping, which is not addressed by this algorithm.

Fig. 12.6 shows several estimated pitch contours from the mixture in Fig. 12.2(d) obtained iteratively as described above. For most time frames, the detected contours well match the reference pitch contours generated from the clean utterance using *Praat* - a standard pitch determination algorithm for clean speech [4].

The above algorithm only tracks one pitch at a frame. When interference also contains a harmonic component, e.g., another utterance, it is probably more helpful to track multiple pitch contours from different sources simultaneously. Wu et al. [54] proposed a robust multipitch tracking algorithm, which works as follows. After a T-F analysis and computing the correlogram, their algorithm selects channels that likely contain signals dominated by harmonic sources. The other channels mostly contain aperiodic sounds and therefore are ignored in subsequent processing. Within each channel, the algorithm treats a peak in the auto-correlation as a pitch hypothesis. Then it integrates periodicity information across the selected channels in order to formulate the conditional probabilities of multiple pitch hypotheses given the periodicity information in these channels. Finally, a continuous hidden Markov model (HMM) is used to model pitch dynamics across successive time frames and the Viterbi algorithm is then used to find optimal pitch contours. The Wu et al. algorithm is illustrated in Fig. 12.7 for pitch tracking of two simultaneous utterances. The algorithm successfully tracks the pitch contours of both utterances at most time frames.

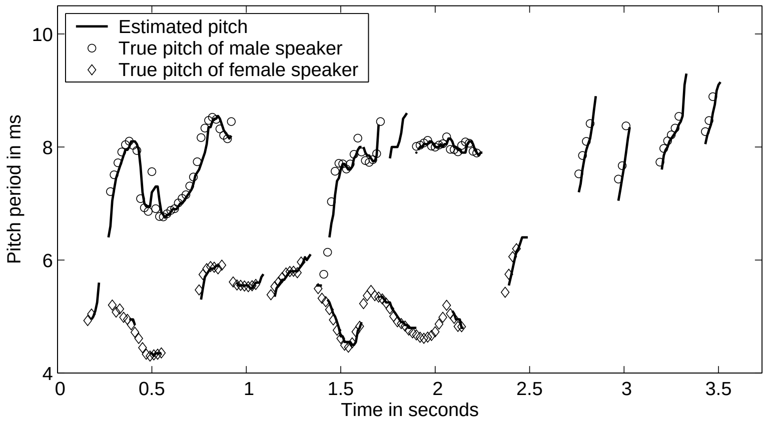

Fig. 12.7. Results of multipitch tracking by the Wu et al. algorithm. The input is the mixture of the utterance in Fig. 12.2(b) and a female utterance: "That noise problem grows more annoying each day." Solid lines indicate estimated target pitch contours. True pitch points of the male utterance are marked by circles, and those of the female utterance are marked by diamonds. For clarity, every other frame is displayed.

12.4 Auditory Segmentation

In addition to the conceptual importance of segmentation in ASA, a segment as a region of T-F units contains more global information of the source that is missing from individual T-F units, such as spectral and temporal envelope. This information could be key for distinguishing sounds from different sources. One may skip the stage of segmentation by grouping individual T-F units directly. However, such grouping based on local information will not be very robust. In our view, auditory segmentation provides a foundation for grouping and is essential for successful CASA.

12.4.1 Segmentation for Voiced Speech

Speech signal lasts for a period of time, within which it has good temporal continuity. Therefore, T-F units neighboring in time tend to originate from the same source. In addition, because the passbands of adjacent channels have significant overlap, a harmonic usually activates a number of adjacent channels, which leads to high cross-channel correlation. Therefore, we perform segmentation by merging T-F units based on temporal continuity and cross-channel correlation [52]. More specifically, only units with sufficiently high cross-channel correlation of correlogram responses are marked, and neighboring marked units are iteratively merged into segments. To account for AM effects of unresolved harmonics, we separately mark and merge high-frequency units on the basis of cross-channel correlation of response envelopes.

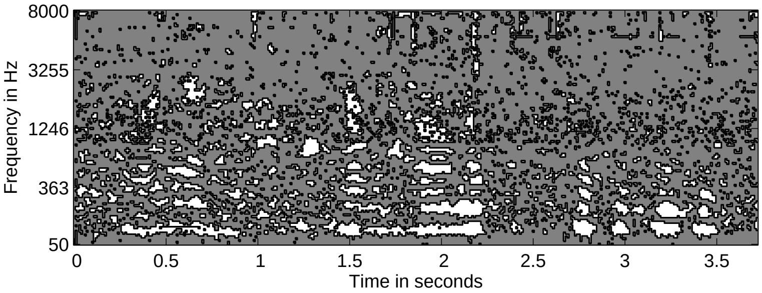

Fig. 12.8. The bounding contours of estimated segments based on cross-channel correlation and temporal continuity. The background is represented by gray.

Fig. 12.8 shows the segments generated in this process for the mixture in Fig. 12.2(d). Compared with Fig. 12.2(e), computed segments cover most T-F regions dominated by voiced speech. In addition, T-F regions dominated by target and interference are well separated into different segments. If desired, very small segments can be easily removed [23]. Note that the correlogram is a periodicity representation, and correlogram-based segmentation therefore is not expected to work well for aperiodic signal, such as unvoiced speech.

12.4.2 Segmentation Based on Onset/Offset Analysis

Unvoiced speech lacks the harmonic structure, and as a result is more difficult to segment. We have proposed a general method for segmentation based on analysis of event onset and offset. This method has three stages: Smoothing, onset/offset detection, and multiscale integration [24], and it works for both voiced and unvoiced speech since onsets and offsets are generic sound properties.

As discussed in Sec. 12.3.3, onsets and offsets correspond to sudden intensity increases and decreases, or the peaks and valleys of the time derivative of the intensity. In smoothing, the intensity is first smoothed over time in order to reduce insignificant fluctuations. We then perform smoothing over frequency to enhance synchronized onsets and offsets across frequency. The degree of smoothing is referred to as the scale [43]. A larger scale leads to smoother intensity.

In the stage of onset/offset detection and matching, our system detects onsets and offsets in each filter channel and merges them into onset and offset fronts if they are sufficiently close. A front corresponds to a boundary along the frequency (vertical) axis in a 2-D cochleagram representation. Individual onset and offset fronts are matched, and a matching pair encloses a segment.

Smoothing with a large scale may blur onsets and offsets of a short acoustic event. Consequently, segmentation may miss short events or combine different events into one segment. On the other hand, smoothing with a small (fine)

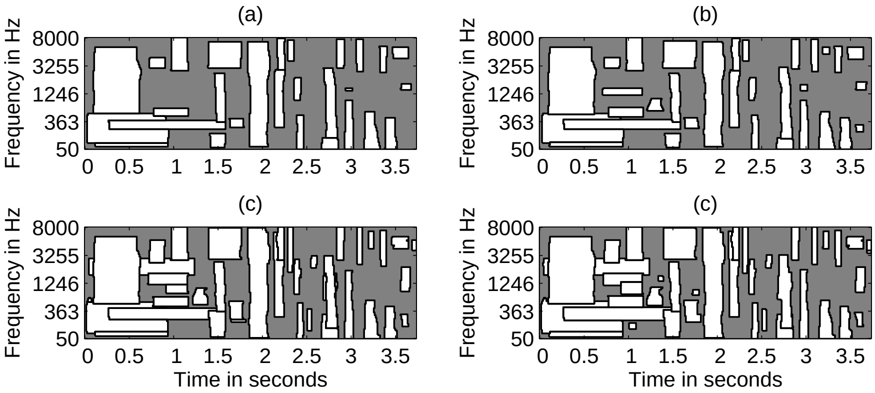

Fig. 12.9. Bounding contours of estimated segments from multiscale analysis of onset and offset. (a) One scale analysis. (b) Two-scale analysis. (c) Three-scale analysis. (d) Four-scale analysis. The input is the mixture shown in Fig. 12.2(d). The background is represented by gray.

scale may not adequately remove insignificant intensity fluctuations. Consequently, segmentation may separate a continuous event into several segments. In general, it is difficult to obtain satisfactory segmentation with a single scale. The multiscale analysis stage is designed to detect and localize different events at appropriate scales. In this stage, we start at a large scale and then gradually move to the finest scale. At each scale, the system generates new segments from within the current background and locates more accurate onset and offset positions for existing segments.

Figs. 12.9(a), 12.9(b), 12.9(c), and 12.9(d) show the bounding contours of obtained segments by integrating 1, 2, 3, and, 4 scales, respectively (see [24] for implementation details). The input is the mixture in Fig. 12.2(d). Comparing it with Fig. 12.2(e), we can see that at the largest scale, the system captures most of the speech events, but misses some small segments. As the system integrates more fine scales, more segments for speech as well as for interference appear.

12.5 Voiced Speech Grouping

To group voiced speech, we use the segments obtained by the simple algorithm described in Sec. 12.4.1. Given pitch contours from the target pitch tracking described in Sec. 12.3.4, we label each T-F unit as target dominant or interference dominant according to target pitch. To label a T-F unit, we first compare the periodicity of its response with the estimated pitch. Specifically, a T-F unit u_{cm} is labeled as target if the correlogram response at the estimated pitch period $\tau_S(m)$ is close to the maximum of the autocorrelation

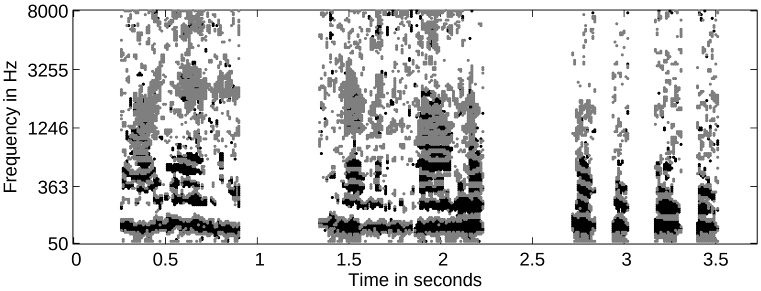

Fig. 12.10. Results of T-F unit labeling for the mixture in Fig. 12.2(d). Black regions: units labeled as target by the periodicity criterion; gray regions: units labeled as target by the AM criterion.

within the plausible pitch range, $\mathit{\Gamma}$:

$$\frac{A_{\mathrm{H}}\big(c, m, \tau_{\mathrm{S}}(m)\big)}{\max\limits_{\tau \in \mathit{\Gamma}} A_{\mathrm{H}}\big(c,\ m, \tau\big)} > \theta_{\mathrm{T}}\,. \tag{12.8}$$

The above criterion, referred to as the *periodicity criterion*, works well for resolved harmonics.

For units responding to multiple harmonics, their responses are amplitude-modulated. We have found that the periodicity criterion does not work well for such units. Observe that the envelope of such a response fluctuates at the f_0 rate of the source. Therefore, we label these T-F units by comparing their AM rates with the estimated pitch. A straightforward way is to check the autocorrelation of response envelopes:

$$\frac{A_{\mathrm{E}}\big(c, m, \tau_{\mathrm{S}}(m)\big)}{\max\limits_{\tau \in \mathit{\Gamma}} A_{\mathrm{E}}\big(c, m, \tau\big)} > \theta_{\mathrm{A}}\,. \tag{12.9}$$

This criterion is referred to as the *AM criterion*.

In practice, we use the periodicity criterion to label T-F units that belong to segments formed on the basis of high cross-channel correlation of filter responses. Such units correspond to resolved harmonics. The remaining units are labeled by the AM criterion.

Fig. 12.10 shows the T-F units labeled as target for the mixture in Fig. 12.2(d). Compared with Fig. 12.2(e), one can see that most units dominated by target voiced speech are correctly labeled. However, some units containing stronger intrusion are also labeled as target speech, especially in the high-frequency range.

With unit labels, we group a segment into the target stream if the acoustic energy corresponding to its T-F units labeled as target exceeds half of the total energy of the segment. Furthermore, significant T-F regions labeled as inference are removed from the target stream. Finally, to group more target

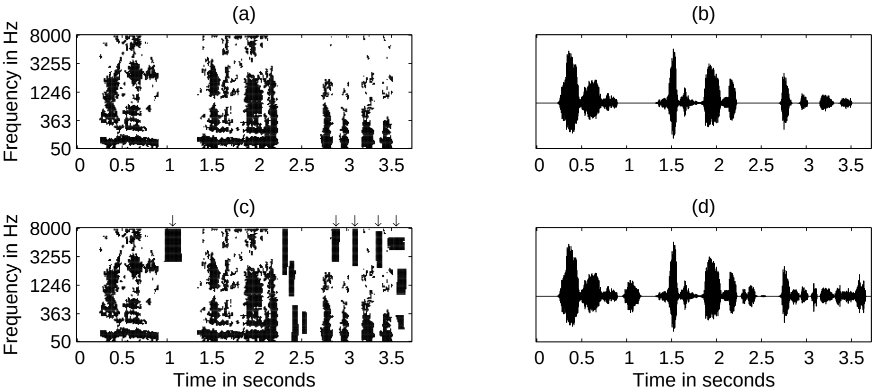

Fig. 12.11. Results of segregation for the mixture in Fig. 12.2(d). (a) Segregated voiced target. (b) The corresponding resynthesized voiced target. (c) Segregated final target. The arrows indicate the segregated fricatives and affricates. (d) Corresponding resynthesized final target.

energy we expand each target segment by iteratively grouping its neighboring units that are labeled as target and do not belong to any segment. When this expansion ends, the system yields a target stream and its background that consists of the remaining T-F units.

Figs. 12.11(a) and 12.11(b) shows the final target stream and the corresponding resynthesized speech for the mixture in Fig. 12.2(d). Compared with Fig. 12.2(e), this stream contains a majority of the T-F units where voiced target speech dominates. In addition, only a small number of units where intrusion dominates are incorrectly included. The segregated speech waveform in Fig. 12.11(b) within voiced speech sections is much more similar to that of the clean speech in Fig. 12.2(b) than the mixture waveform in Fig. 12.2(d).

The performance of the system on voiced speech segregation has been evaluated using a corpus of 100 mixtures composed of 10 voiced utterances mixed with 10 intrusions collected by Cooke [13]. This corpus has been used to test previous CASA systems [13, 7, 16, 52, 15]. The intrusions have a considerable variety; specifically they are described in Tab. 12.1.

As discussed in Sec. 12.2, our computational goal is to estimate the ideal binary mask. Therefore, our evaluation compares the segregated speech, $\hat{s}(n)$, against the speech waveform resynthesized from the ideal binary mask, $s(n)$. Let $e_1(n)$ denote the signal present in $s(n)$ but missing from $\hat{s}(n)$, and $e_2(n)$ the signal present in $\hat{s}(n)$ but missing from $s(n)$. Then, we measure the percentage of energy loss, P_{EL}, and the percentage of noise residue, P_{NR}:

$$P_{\mathrm{EL}} = \sum_n e_1^2(n) \Bigg/ \sum_n s^2(n) \,, \tag{12.10}$$

Table 12.1. Types of intrusions.

Intrusion	Description
N0	1kHz pure tone
N1	white noise
N2	noise bursts
N3	"cocktail party" noise
N4	rock music
N5	siren
N6	trill telephone
N7	female speech
N8	male speech
N9	female speech

$$P_{\mathrm{NR}} = \sum_n e_2^2(n) \bigg/ \sum_n \hat{s}^2(n)\,. \tag{12.11}$$

P_{EL} indicates the percentage of target speech excluded from segregated speech, and P_{NR} the percentage of intrusion included. They provide complementary error measures of a segregation system and a successful system needs to achieve low errors in both measures.

The results from our model are shown in Tab. 12.2. Each value in the table represents the average result of one intrusion with 10 voiced utterances, and a further average across all intrusions is also shown. On average, our system retains 96.28% of target speech energy, and the percentage of noise residue is kept at 2.81%. The percentage of noise residue for the original mixtures is 36.05%, also shown in the table; energy loss is obviously zero for the original mixtures. As indicated by the table, our model achieves very good performance across the noise types. In particular, the errors measured by P_{EL} and P_{NR} are balanced in our system.

Since our model applies different mechanisms to segregate resolved and unresolved harmonics, it is instructive to present the performance in the high-frequency range separately. For this purpose, we calculate the percentages of energy loss and noise residue for only the filter channels with center frequencies greater than 1 kHz, denoted by $P_{\mathrm{EL}}^{\mathrm{H}}$ and $P_{\mathrm{NR}}^{\mathrm{H}}$, respectively. Note that for the evaluation corpus, target harmonics in the frequency range above 1 kHz are generally unresolved. The corresponding results are shown in Tab. 12.2. Most of the voiced energy in the high-frequency range is recovered and not much interference is included. The performance in high-frequency range is not as good as that in the low-frequency range since intrusions are relatively much stronger in the high-frequency range, which is clear from the average noise residue of the original mixtures and that in the high-frequency range.

To compare waveforms directly we can measure SNR in decibels:

Table 12.2. P_{EL} and P_{NR} for segregation of voiced speech.

Intrusion	**Segregated target**				**Mixture**	
	$P_{\mathrm{EL}}(\%)$	$P_{\mathrm{EL}}^{\mathrm{H}}(\%)$	$P_{\mathrm{NR}}(\%)$	$P_{\mathrm{NR}}^{\mathrm{H}}(\%)$	$P_{\mathrm{NR}}(\%)$	$P_{\mathrm{NR}}^{\mathrm{H}}(\%)$
N0	1.47	14.97	0.05	0.52	67.76	96.82
N1	4.61	32.48	3.78	61.00	57.16	96.00
N2	1.01	8.18	0.42	7.98	5.04	44.02
N3	4.04	12.90	2.14	6.44	18.15	42.57
N4	2.81	21.42	3.58	43.28	27.17	81.31
N5	1.32	7.47	0.06	0.46	78.84	97.90
N6	0.95	8.99	0.94	16.27	39.24	91.26
N7	2.01	9.76	2.25	8.68	16.68	43.49
N8	1.16	8.59	0.65	4.32	7.37	31.07
N9	17.80	19.25	14.22	5.47	43.09	27.72
Average	3.72	14.40	2.81	15.44	36.05	65.22

$$SNR = 10 \log_{10} \frac{\sum\limits_n s^2(n)}{\sum\limits_n \left[s(n) - \hat{s}(n)\right]^2} \, . \tag{12.12}$$

The SNR for each intrusion averaged across 10 target utterances is shown in Fig. 12.12, together with the SNR of the original mixtures and the results from the Wang-Brown system [52], whose performance is representative of previous CASA systems, and a spectral subtraction method [5], a standard method for speech enhancement. Our system shows substantial improvements. In particular, it yields a 12.1 dB gain on average over the original mixtures, a 5.8 dB gain over the Wang-Brown model, and a 7.0 dB gain over spectral subtraction.

We point out that, although the above algorithm for voiced speech segregation is similar to that presented in [23], it is simplified a good deal. The guiding principle for the algorithm presented in this chapter is to simplify that in [23] as much as possible without sacrificing the segregation performance. Also the delay compensation for gammatone filters discussed in Sec. 12.3.1 is not implemented in [23]. Indeed, the SNR performance for the simplified version is even slightly better than that in [23]. For completeness, we give the entire algorithm in the Appendix along with a few further notes.

12.6 Unvoiced Speech Grouping

Unvoiced speech lacks the periodicity feature, which plays the primary role in voiced speech segregation, and segregation of unvoiced speech is particularly challenging. Unvoiced speech in English contains three categories of consonants: Stops, fricatives, and affricates [30]. Stops consist of /t/, /d/, /p/, /b/, /k/, and /g/, and fricatives consist of /s/, /z/, /f/, /v/, /θ/, /ð/, /ʃ/, /ʒ/, and

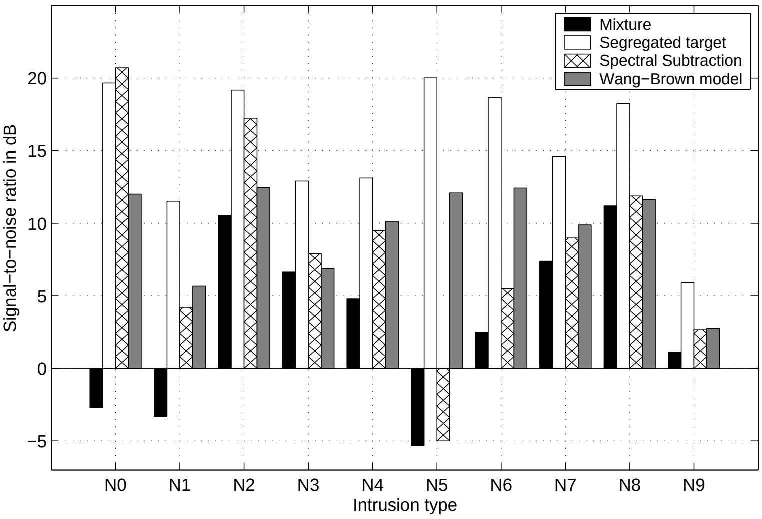

Fig. 12.12. Signal-to-noise ratio (SNR) results against the ideal binary mask for segregated speech and original mixtures. White bars show the results from our system, gray bars those from the Wang-Brown system, cross bars those from a spectral subtraction method, and black bars those of original mixtures.

/h/. There are two affricates, /tʃ/ and /dʒ/, each of which is a stop followed by a fricative. Although about half of these consonants are phonetically voiced, their acoustic realizations often contain weak voicing [50], and they cannot be reliably segregated with pitch-based analysis. Hence all these consonants are treated in this section. As stated in Sec. 12.2, here we only deal with nonspeech interference. Because of the similarity between fricatives and affricates, we consider them together. In this section, we first describe segregation of stop consonants and then segregation of fricatives and affricates.

12.6.1 Segregation of Stop Consonants

A stop consonant starts with a closure corresponding to the stop of airflow in the vocal tract, followed by a burst corresponding to a sudden release of airflow. The closure contains little energy and is usually masked by interference. The focus here is to segregate stop bursts.

In a previous study, we have proposed to segregate stop consonants in two steps: Stop detection and stop grouping [22]. In the first step, onset detection is performed in each frequency channel, and onset fronts are formed by connecting close onsets at neighboring channels. We distinguish onset fronts belonging to stop consonants from others via featured-based classification. Stop bursts are characterized by the following features: Spectral envelope, intensity, duration, and formant transition (see [1] for example). However, the

Table 12.3. P_{EL} and P_{NR} for stop consonants.

Overall SNR (dB)	$P_{\text{EL}}(\%)$	$P_{\text{NR}}(\%)$
0	84.79	9.62
10	70.68	2.81
20	41.56	0.81
30	28.01	0.04

formant transition from a stop to its neighboring voiced phoneme is very difficult to obtain; moreover, it is closely related to the spectrum. Therefore we use the following features for classification: Spectral envelope, intensity, and duration.

Stop consonants are grouped based on onset synchrony. Specifically, for each detected stop, the frequency channels that contain onsets synchronous with the onset of the stop burst are grouped together. The temporal boundary within each such channel is determined as from the minimum filter response immediately before the burst duration to the minimum point immediately after the burst. This pair of minima approximately marks the onset and the offset of the stop for the filter channel. The T-F units within this interval are hence labeled as belonging to the stop consonant.

The above method has been tested with 10 utterances from the TIMIT database mixed with the following 10 interference: White noise, pink noise, airplane noise, car noise, factory noise, noise burst, clicks, bar noise, fireworks, and rain. Average P_{EL} and P_{NR} for stop consonants at different SNR levels are shown in Tab. 12.3. The system performs well when SNR is relatively high. As SNR decreases, P_{EL} increases significantly while P_{NR} remains relatively low.

12.6.2 Grouping of Fricatives and Affricates

We group fricatives and affricatives with the segments obtained by the segmentation algorithm described in Sec. 12.4.2. Because fricatives and affricates are relatively steady acoustically [50], most T-F units dominated by these consonants are well organized into obtained segments. The task here is to distinguish these segments from those corresponding to interference. This is performed in two steps [25]. First, we remove those segments dominated by non-fricative and non-affricate sounds within voiced sections. Then we apply a Bayesian classifier to determine whether each remaining segment belongs to a fricative, an affricate, or interference.

The motivation of the first step is to take advantage of segregated voiced speech. In the segmentation stage described in Sec. 12.4.2, obtained segments containing significant portions of fricatives and affricates tend to contain little signal from other phonemes or interference. Therefore, segments overlapping significantly with non-fricative and non-affricate sounds are removed. To identify these segments, our system first uses the segregated voiced speech to determine time frames containing phonemes other than fricatives and affricates as follows.

Let H_0 be the hypothesis that a T-F region is dominated by interference, $H_{1,k}$ a T-F region dominated by a fricative or an affricate, indexed by k, and $H_{2,l}$ a T-F region dominated by another phoneme, indexed by l. Let $X(m)$ be the power spectrum of the input mixture at frame m, and $X_\mathrm{S}(m)$ be the corresponding power spectrum within segregated target stream. Frame m is labeled as non-fricative and non-affricate if

$$\max_k P\big(H_{1,k}\big|X_\mathrm{S}(m)\big) < \max_l P\big(H_{2,l}\big|X_\mathrm{S}(m)\big)\,. \tag{12.13}$$

By applying the Bayesian rule, we have

$$\max_k \Big[p\big(X_\mathrm{S}(m)\big|H_{1,k}\big)\,P\big(H_{1,k}\big)\Big] < \max_l \Big[p\big(X_\mathrm{S}(m)\big|H_{2,l}\big)\,P(H_{2,l})\Big]\,. \tag{12.14}$$

Note that frames not occupied by the segregated target are not considered. The segments whose energy is dominated by such frames are removed.

For each remaining segment, which lasts from frame m_1 to m_2, let $Y(m)$ be the power spectrum within the segment at frame m, and

$$\boldsymbol{Y} = \big[Y(m_1), Y(m_1+1), \ldots, Y(m_2)\big]\,. \tag{12.15}$$

This segment is classified as dominated by a fricative or an affricate if:

$$\max_k \Big[p\big(\boldsymbol{Y}\big|H_{1,k}\big)\,P\big(H_{1,k}\big)\Big] > p\big(\boldsymbol{Y}\big|H_0\big)\,P\big(H_0\big)\,. \tag{12.16}$$

Because segments have varied sizes, the complexity for computing $p(\boldsymbol{Y}|H_{1,k})$ and $p(\boldsymbol{Y}|H_0)$ directly is very high. Fortunately, we find that, by considering only the dependence between two consecutive frames, a good estimate of $p(\boldsymbol{Y}|H_0)$ can be obtained,

$$p\big(\boldsymbol{Y}\big|H_0\big) = p\big(Y(m_1)\big|H_0\big)\prod_{m=m_1}^{m_2-1} p\big(Y(m+1)\big|Y(m),\,H_0\big)\,. \tag{12.17}$$

This observation holds for $p(\boldsymbol{Y}|H_{1,k})$ also. Then Eq. 12.16 becomes

$$\begin{aligned}&\max_k \Big[p\big(Y(m_1)\big|H_{1,k}\big)\,P\big(H_{1,k}\big)\prod_{m=m_1}^{m_2-1} p\big(Y(m+1)\big|Y(m),\,H_{1,k}\big)\Big]\\ &\quad > p\big(Y(m_1)\big|H_0\big)\,P\big(H_0\big)\prod_{m=m_1}^{m_2-1} p\big(Y(m+1)\big|Y(m),\,H_0\big)\,.\end{aligned} \tag{12.18}$$

In Eq. 12.18, segment duration is implicitly given. To emphasize the contribution of duration in classification, we insert duration D as an additional feature into Eq. 12.18:

$$\begin{aligned}\max_k \Big[p\big(Y(m_1),\, D \big| H_{1,k}\big)\, P\big(H_{1,k}\big) \prod_{m=m_1}^{m_2-1} p\big(Y(m+1),\, D \big| Y(m),\, H_{1,k}\big) \Big] \\ > p\big(Y(m_1),\, D \big| H_0\big)\, P\big(H_0\big) \prod_{m=m_1}^{m_2-1} p\big(Y(m+1),\, D \big| Y(m),\, H_0\big)\,, \end{aligned} \tag{12.19}$$

so that the contributions from spectrum and duration are well balanced.

We use the two features of spectrum (including the spectral envelope and intensity) and duration for the classification task in both of the steps. The formant transition is another feature for identifying fricatives and affricates. As discussed in Sec. 12.6.1, the formant transition is partly captured by the spectrum. In addition, it is very difficult to extract. Therefore, it is not utilized here.

The prior distributions and probabilities required for calculating Eq. 12.14 and Eq. 12.19 are obtained from training using the training part of the TIMIT database and 90 environmental intrusions, including crowd noise, traffic noise, and wind, etc. A Gaussian mixture model with 8 components and a full covariance matrix for each mixture is used for training the probability density function for all the spectral features and duration. Then in calculating Eq. 12.14 and Eq. 12.19, we use marginal distribution since only a subset of spectral features is included in the formula.

All the segments identified as dominated by fricatives or affricates are added to the segregated voiced target. As an illustration, Figs. 12.11(c) and 12.11(d) show the final target stream and the corresponding resynthesized speech for the mixture in Fig. 12.2(d). The target utterance, "*H*er right *h*and ache*s* whene*v*er *th*e barometric pre*ss*ure *ch*an*g*e*s*" contains 7 fricatives and 2 affricates, italicized in the sentence. Among them, /h/ in "hand", /v/ in "whenever", and /ð/ in "the" are mainly voiced and portions of their energy are recovered in voiced speech segregation (see Fig. 12.11(a)). /h/ in "her" is mostly masked by the intrusion, hence not recoverable. The remaining 5 are successfully segregated by the system, as indicated by the arrows in Fig. 12.11(c). At the same time, some intrusion-dominated T-F regions are also included in the segregated target.

The performance of fricative and affricate segregation is systematically evaluated with 20 utterances from the testing part of the TIMIT database, mixed with 10 intrusions at different SNR levels. The intrusions are white noise, electrical fan, rooster crowing and clock alarm, traffic noise, crowd noise in playground, crowd noise with music, crowd noise with clapping, bird chirping and water flow, wind, and rain.

Tab. 12.4 shows the average P_{EL} and P_{NR} for segregation of fricatives and affricates. As shown in the table, our system extracts about 70% of the fricative and affricate energy from the mixture under different SNR situations. On the other hand, it retains certain amounts of interference, which are much less than those included in the original mixture. Our system performs significantly better than a spectral subtraction method, especially in low SNR situations [25].

Table 12.4. P_{EL} and P_{NR} for fricatives and affricates.

Overall SNR (dB)	**Segregated target**		**Mixture**
	$P_{\mathrm{EL}}(\%)$	$P_{\mathrm{NR}}(\%)$	$P_{\mathrm{NR}}(\%)$
0	33.48	35.11	82.17
5	32.39	21.19	61.38
10	29.39	8.47	36.05
15	29.60	5.34	16.39
20	29.88	3.30	6.21

12.7 Concluding Remarks

We should point out that our approach is primarily feature-based. The features used by the system, such as periodicity, AM, and onset, are general properties. Our system does not employ specific prior knowledge of target or interference, except in unvoiced speech grouping where we perform phonetic classification. Prior knowledge helps human ASA in the form of schema-based grouping [6]. Schema-based organization has been emphasized by Ellis [16], and is a subject of several recent studies. Roweis trained HMMs to separate mixtures from two speakers [45]. Barker et al. coupled segmentation with explicit speech models [2]. Srinivasan and Wang used word models to restore phonemes that are masked by interference [49]. These model-based approaches should help to improve the performance of a feature-based system.

A natural speech utterance contains silent gaps and other sections masked by interference. In practice, one needs to group the utterance across such time intervals. This is the problem of sequential grouping, which is not addressed in this chapter. One way of grouping segments across time uses speech recognition in a top-down manner [2]. Recently, Shao and Wang proposed to perform sequential grouping [47] using trained speaker models. Such methods can be integrated with simultaneous grouping addressed in this chapter. Room reverberation is another important issue that must be addressed before speech segregation systems can be deployed in real world environments (see [41] for a recent study on pitch-based segregation of reverberant speech).

To conclude, we have described a CASA approach to monaural speech segregation. Our system segregates voiced speech based on periodicity and AM as well as temporal continuity. Unvoiced speech is segregated via onset/offset analysis and feature-based classification. Evaluation results show that the system performs well on both voiced and unvoiced speech. Note that unvoiced speech is particularly challenging for monaural speech segregation, and our research is the first systematic study on separating unvoiced speech.

Appendix: Voiced Speech Segregation Algorithm

In this appendix, we provide the complete algorithm for voiced speech segregation along with several notes. To facilitate the reader's use of this algorithm, we also post the C++ code for the algorithm on the website (http://www.cse.ohio-state.edu/pnl/software.html). See text for notations. The parameter values used in our implementation are: $\theta_{\mathrm{C}} = 0.99$, $\theta_{\mathrm{P}} = 0.95$, $\theta_{\mathrm{T}} = 0.85$, and $\theta_{\mathrm{A}} = 0.7$. The plausible pitch period range, $\mathit{\Gamma}$, is [2 ms, 12.5 ms]. The algorithm is given below.

1. **Cochlear filtering.** A bank of 128 gammatone filters centered from 80 Hz to 5000 Hz is used.

2. **Auditory nerve transduction.** The Meddis model is used.

3. **Feature extraction.** The following features are extracted: Correlogram, envelope correlogram, cross-channel correlation, and dominant pitch. The envelope is obtained through half-wave rectification and bandpass filtering with the passband from 50 Hz to 550 Hz.

4. **Segmentation**

4.1. Mark two adjacent T-F units, u_{cm} and $u_{c+1,m}$, according to their cross-channel correlation:

4.1.1. If $C_{\mathrm{H}}(c, m) > \theta_{\mathrm{C}}$, both units are marked as 1.

4.1.2. Else if $C_{\mathrm{E}}(c, m) > \theta_{\mathrm{C}}$ and the center frequency of channel c is above 1 kHz, both units are marked as 2.

4.2. Neighboring T-F units with the same mark are merged into segments. Two types of segments are obtained, type 1 and type 2, according to their marks. Two units are considered neighbors if they share the same channel and appear in consecutive time frames, or if they share the same frame and appear in adjacent filter channels. Note that there are unmarked units.

5. **Target pitch tracking**

5.1. Initial grouping. Only type-1 segments are considered.

5.1.1. u_{cm} is labeled as the dominant source if

$$\frac{A_{\mathrm{H}}\big(c, m, \tau_S(m)\big)}{\max\limits_{\tau \in \mathit{\Gamma}} A_{\mathrm{H}}\big(c, m, \tau\big)} > \theta_{\mathrm{P}} .$$

$\tau_S(m)$ initially indicates the dominant pitch period at frame m.

5.1.2. At a frame of a segment, the segment is labeled as the dominant source if its T-F units labeled as the dominant source contain more than half of the total energy of the segment at the frame; otherwise, it is labeled as the background.

5.1.3. Find a seed segment that has the largest number of frames labeled as the dominant source.

5.1.4. Determine whether a segment agrees with the seed segment. A segment agrees with the seed segment if they share the same label (either dominant source or background) for more than 2/3 of their overlapping frames. All the segments agreeing with the seed segment form an initial estimate of the target stream, S_0.

5.2. Estimate the target pitch contour from S_0 for every frame of the seed segment. For each such frame, m, the estimated target pitch period, $\tau_S(m)$, is the lag corresponding to the maximum of $\sum_{c, u_{cm} \in S_0} A_{\mathrm{H}}(c, m, \tau)$ in Γ.

5.3. Label individual T-F units and check the reliability of the estimated pitch against the *consistency constraint*: A reliable target pitch is consistent with the periodicity of S_0.

5.3.1. Label a T-F unit at frame m with an estimated pitch as target if

$$\frac{A_{\mathrm{H}}\big(c, m, \tau_S(m)\big)}{\max\limits_{\tau \in \Gamma} A_{\mathrm{H}}\big(c, m, \tau\big)} > \theta_{\mathrm{P}} .$$

Otherwise, label it interference.

5.3.2. If less than half of the T-F units of S_0 at frame m are labeled as target, the estimated pitch, $\tau_S(m)$, is considered inconsistent and all the T-F units of frame m are labeled as interference.

5.4. Re-estimate target stream with labeled T-F units. A segment is labeled as target if its T-F units labeled as target contain more than half of its total energy. All the segments labeled as target form a new estimate of target, S_1.

5.5. Estimate target pitch for all the frames of S_1 as done in Step 5.2. Label individual T-F units and check the consistency of the estimated pitch as done in Step 5.3.

5.6. Pitch interpolation for frames with unreliable pitch:

5.6.1. Consistent pitch points in consecutive frames are connected to form a set of smooth contours. A smooth contour is the one where consecutive frames on the contour satisfy the *smoothness constraint*: The pitch contour of speech changes slowly. Specifically, the change from a pitch period to the one at the next frame is considered smooth if the change is less than 20% of both pitch periods.

5.6.2. Find the longest smooth contour and denote it the seed contour.

5.6.3. Re-estimate the pitch periods for the frames before the seed contour. Set m to the first frame of the seed contour. Iterate until m is the first frame of S_1:

i. Denote the current frame, m, as a reliable frame (i.e. it has a reliable pitch estimate) and denote c as a selected channel if $u_{cm} \in S_1$ and is labeled as target.

ii. Decrease m by 1.

iii. Check if $\tau_{\mathrm{S}}(m)$ satisfies both the consistency and the smoothness constraints. If yes, go directly to Step 5.6.3.i.

iv. Summate the autocorrelations of u_{cm}'s at frame m where $u_{cm} \in S_1$ and c is a selected channel of the nearest reliable frame. Replace $\tau_{\mathrm{S}}(m)$ by the lag corresponding to the maximum of the summation in the range $[0.65\tau_{\mathrm{R}},\ 1.55\tau_{\mathrm{R}}]$, where τ_{R} indicates the estimated pitch period at the nearest reliable frame.

v. Check if the new $\tau_{\mathrm{S}}(m)$ satisfies the smoothness constraint. If not, $\tau_{\mathrm{S}}(m)$ is considered unreliable, and then go directly to Step 5.6.3.ii.

5.6.4. Re-estimate the pitch periods for the frames after the seed contour in a symmetric way, until the last frame of S_1.

5.6.5. For any interval of unreliable pitch estimates between two intervals of reliable estimates, the pitch periods within this interval are obtained by linear interpolation from the last frame of the preceding reliable interval and the first frame of the succeeding one.

6. **T-F unit labeling**

6.1. For unit u_{cm} belonging to a type-1 segment, label it as target if

$$\frac{A_{\mathrm{H}}\big(c, m, \tau_{\mathrm{S}}(m)\big)}{\max\limits_{\tau \in \mathit{\Gamma}} A_{\mathrm{H}}\big(c, m, \tau\big)} > \theta_{\mathrm{T}}\,.$$

Otherwise, label it as interference.

6.2. For a remaining unit, u_{cm}, label it as target if

$$\frac{A_{\mathrm{E}}\big(c, m, \tau_{\mathrm{S}}(m)\big)}{\max\limits_{\tau \in \mathit{\Gamma}} A_{\mathrm{E}}\big(c, m, \tau\big)} > \theta_{\mathrm{A}}\,.$$

Otherwise, label it as interference.

7. **Grouping**

7.1. A segment is labeled as target if its T-F units labeled as target contain more than half of its total energy. These segments form S_2.

7.2. In S_2, find all the contiguous T-F regions that are all labeled as interference, and remove those regions longer than 40 ms.

7.3. Expand S_2 by iteratively grouping neighboring unmarked T-F units that are labeled as target.

The resulting S_2 represents the segregated target speech by the algorithm. A few further notes are in order. Regarding Step 2 – the modeling of the auditory nerve transduction – we find that the performance without the step is similar for all intrusions except N9, a female utterance. Step 2 helps the system to obtain a better target pitch estimate with the N9 intrusion.

Note also that the algorithm segregates only one continuous section of voiced speech since the pitch determination algorithm provides one pitch contour. If multiple pitch contours are given, one can easily use the given contours instead of Step 5. As discussed in Sec. 12.3.4, we can also apply Step 5 iteratively to estimate multiple pitch contours. However, there is no guarantee that a pitch contour generated this way corresponds to target speech. As mentioned in Sec. 12.7, to determine whether a pitch contour is a target contour is the task of sequential grouping, not addressed here. Step 5.6 in the above algorithm performs pitch interpolation and is relatively complicated. A simpler way is to perform linear interpolation between smooth contours obtained in Step 5.6.1. However, we find this simple method does not work as well for two reasons. First, our tracking algorithm attempts to re-estimate unreliable pitch points from selected frequency channels at the nearest reliable frame, an instance of applying temporal continuity. Second, some smooth contours are inaccurate – e.g. reflecting doubles of pitch frequencies – and when this happens, the smoothness of the overall pitch contour tends to be violated. The tracking algorithm from a seed contour guarantees the smoothness of an overall pitch contour.

Acknowledgments

The preparation of this chapter was supported in part by an AFOSR grant (FA9550-04-01-0117) and an AFRL grant (FA8750-04-1-0093).

References

1. A.M.A. Ali, J. Van der Spiegel: Acoustic-phonetic features for the automatic classification of stop consonants, *IEEE Trans. Speech Audio Process.,* **9**, 833–841, 2001.
2. J.P. Barker, M.P. Cooke, D.P.W. Ellis: Decoding speech in the presence of other sources, *Speech Comm.,* **45**, 5–25, 2005.
3. J. Bird, C.J. Darwin: Effects of a difference in fundamental frequency inseparating two sentences, in A.R. Palmer, A. Rees, A.Q. Summerfield, R. Meddis (eds.), *Psychophysical and Physiological Advances in Hearing,* London, UK: Whurr, 263–269, 1998.

4. P. Boersma, D. Weenink: *Praat: Doing Phonetics by Computer,* Version 4.2.31, http://www.fon.hum.uva.nl/praat/, 2004.
5. S.F. Boll: Suppression of acoustic noise in speech using spectral subtraction, *IEEE Trans. Acoust. Speech Signal Process.,* **27**, 113–120, 1979.
6. A.S. Bregman: *Auditory Scene Analysis,* Cambridge, MA, USA: MIT Press, 1990.
7. G.J. Brown, M.P. Cooke: Computational auditory scene analysis, *Comput. Speech and Language,* **8**, 297–336, 1994.
8. G.J. Brown, D.L. Wang: Separation of speech by computational auditory scene analysis, J. Benesty, S. Makino, J. Chen (eds.), *Speech Enhancement,* Berlin, Germany: Springer, 371–402, 2005.
9. D.S. Brungart, P.S. Chang, B.D. Simpson, D.L. Wang: Isolating the energetic component of speech-on-speech masking with an ideal binary mask, *Submitted for journal publication,* 2005.
10. J. Canny: A computational approach to edge detection, *IEEE Trans. Pattern Analysis and Machine Intelligence,* **8,** 679–698, 1986.
11. R.P. Carlyon, T.M. Shackleton: Comparing the fundamental frequencies of resolved and unresolved harmonics: evidence for two pitch mechanisms? *J. Acoust. Soc. Am.,* **95**, 3541–3554, 1994.
12. P.S. Chang: *Exploration of Behavioral, Physiological, and Computational Approaches to Auditory Scene Analysis,* M.S. Thesis, The Ohio State University Dept. Comput. Sci. & Eng., 2004 (available at http://www.cse.ohio-state.edu/pnl/theses).
13. M.P. Cooke: *Modelling Auditory Processing and Organisation,* Cambridge, UK: Cambridge University Press, 1993.
14. M.P. Cooke, P. Green, L. Josifovski, A. Vizinho: Robust automatic speech recognition with missing and unreliable acoustic data, *Speech Comm.,* **34**, 267–285, 2001.
15. L.A. Drake: *Sound Source Separation via Computational Auditory Scene Analysis (CASA) – Enhanced Beamforming,* Ph.D. Dissertation, Northwestern University Dept. Elec. Eng., 2001.
16. D.P.W. Ellis: *Prediction-driven Computational Auditory Scene Analysis,* Ph.D. Dissertation, MIT Dept. Elec. Eng. & Comput. Sci., 1996.
17. Y. Ephraim, H.L. van Trees: A signal subspace approach for speech enhancement, *IEEE Trans. Speech Audio Process.,* **3**, 251–266, 1995.
18. J. Garofolo, L. Lamel, et al.: Darpa TIMIT acoustic-phonetic continuous speech corpus, *NISTIR 4930,* 1993.
19. H. Helmholtz: *On the Sensation of Tone,* 2nd English ed., New York, NY, USA: Dover Publishers, 1863.
20. J. Holdsworth, I. Nimmo-Smith, R.D. Patterson, P. Rice: Implementing a gammatone filter bank, *MRC Applied Psych. Unit,* 1988.
21. G. Hu, D.L. Wang: Speech segregation based on pitch tracking and amplitude modulation, *Proc. WASPAA '01*, 79–82, New Paltz, New York, USA, 2001.
22. G. Hu, D.L. Wang: Separation of stop consonants, *Proc. ICASSP '03,* **2**, 749–752, 2003.
23. G. Hu, D.L. Wang: Monaural speech segregation based on pitch tracking and amplitude modulation, *IEEE Trans. Neural Net.,* **15**, 1135–1150, 2004.
24. G. Hu, D.L. Wang: Auditory segmentation based on event detection, *Proc. ISCA Tutorial and Research Workshop on Stat. & Percept. Audio Process.,* 2004.

25. G. Hu, D.L. Wang: Separation of fricatives and affricates, *Proc. ICASSP '05*, **1**, 1101–1104, Philadelphia, PA, USA, 2005.
26. A. Hyvärinen, J. Karhunen, E. Oja: *Independent Component Analysis,* New York, NY, USA: Wiley, 2001.
27. ISO: *Normal Equal-loudness Level Contours for Pure Tones under Free-field Listening Conditions (ISO 226),* International standards organization.
28. J. Jensen, J.H.L. Hansen: Speech enhancement using a constrained iterative sinusoidal model, *IEEE Trans. Speech Audio Process.,* **9**, 731–740, 2001.
29. H. Krim, M. Viberg: Two decades of array signal processing research: The parametric approach, *IEEE Signal Process. Mag.,* **13**, 67–94, 1996.
30. P. Ladefoged: *Vowels and Consonants,* Oxford, UK: Blackwell, 2001.
31. J.C.R. Licklider: A duplex theory of pitch perception, *Experientia,* **7**, 128–134, 1951.
32. D. Marr: *Vision,* New York, NY, USA: Freeman, 1982.
33. R. Meddis: Simulation of auditory-neural transduction: Further studies, *J. Acoust. Soc. Am.,* **83**, 1056–1063, 1988.
34. R. Meddis, M. Hewitt: Modelling the identification of concurrent vowels with different fundamental frequencies, *J. Acoust. Soc. Am.,* **91**, 233–245, 1992.
35. B.C.J. Moore: *An Introduction to the Psychology of Hearing,* 5th ed., San Diego, CA, USA: Academic Press, 2003.
36. R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice: An efficient auditory filterbank based on the gammatone function, *MRC Applied Psych. Unit. 2341,* 1988.
37. J.O. Pickles: *An Introduction to the Physiology of Hearing,* 2nd ed., London, UK: Academic Press, 1988.
38. R. Plomp: The Ear as a Frequency Analyzer, *J. Acoust. Soc. Am.,* **36**, 1628–1636, 1964.
39. R. Plomp: *The Intelligent Ear,* Mahwah, NJ, USA: Lawrence Erlbaum Associates, 2002.
40. R. Plomp, A.M. Mimpen: The ear as a frequency analyzer II, *J. Acoust. Soc. Am.,* **43**, 764–767, 1968.
41. N. Roman, D.L. Wang: A pitch-based model for separation of reverberant speech, *Proc. INTERSPEECH '05,* 2109–2112, Lisbon, Portugal, 2005.
42. N. Roman, D.L. Wang, G.J. Brown: Speech segregation based on sound localization, *J. Acoust. Soc. Am.,* **114**, 2236–2252, 2003.
43. B.H. Romeny, L. Florack, J. Koenderink, M. Viergever (eds.): *Scale-space Theory in Computer Vision,* Berlin, Germany: Springer, 1997.
44. D.F. Rosenthald, H.G. Okuno (eds.): *Computational Auditory Scene Analysis,* Mahwah, NJ: Lawrence Erlbaum Associates, 1998.
45. S.T. Roweis: One microphone source separation, *Proceedings of the Annual Neural Information Processing Systems (NIPS 2000) Conference,* 2001.
46. H. Sameti, H. Sheikhzadeh, L. Deng, R.L. Brennan: HMM-based strategies for enhancement of speech signals embedded in nonstationary noise, *IEEE Trans. Speech Audio Process.,* **6**, 445–455, 1998.
47. Y. Shao, D.L. Wang: Model-based sequential organization in cochannel speech, *IEEE Trans. Speech Audio Process.,* in press, 2005.
48. M. Slaney, R.F. Lyons: A perceptual pitch detector, *Proc. ICASSP '90*, **1**, 357–360, Albuquerque, NM, USA, 1990.
49. S. Srinivasan, D.L. Wang: A schema-based model for phonemic restoration, *Speech Comm.,* **45**, 63–87, 2005.

50. K.N. Stevens: *Acoustic Phonetics,* Cambridge, MA, USA: MIT Press, 1998.
51. D.L. Wang: On ideal binary mask as the computational goal of auditory scene analysis, P. Divenyi (ed.), *Speech Separation by Humans and Machines,* Norwell, MA, USA: Kluwer, 181–197, 2005.
52. D.L. Wang, G.J. Brown: Separation of speech from interfering sounds based on oscillatory correlation, *IEEE Trans. Neural Net.,* **10**, 684–697, 1999.
53. M. Weintraub: *A Theory and Computational Model of Auditory Monaural Sound Separation,* Ph.D. Dissertation, Stanford University Dept. Elec. Eng., 1985.
54. M. Wu, D.L. Wang, G.J. Brown: A multipitch tracking algorithm for noisy speech, *IEEE Trans. Speech Audio Process.,* **11**, 229–241, 2003.