

16 Separation of Speech by Computational Auditory Scene Analysis

Guy J. Brown¹ and DeLiang Wang²

¹ Department of Computer Science, University of Sheffield,
Regent Court, 211 Portobello Street, Sheffield S1 4DP, United Kingdom
E-mail: g.brown@dcs.shef.ac.uk

² Department of Computer Science & Engineering and Center for Cognitive
Science, The Ohio State University, 2015 Neil Avenue, Columbus, OH
43210-1277, USA
E-mail: dwang@cse.ohio-state.edu

Abstract. The term *auditory scene analysis (ASA)* refers to the ability of human listeners to form perceptual representations of the constituent sources in an acoustic mixture, as in the well-known ‘cocktail party’ effect. Accordingly, *computational auditory scene analysis (CASA)* is the field of study which attempts to replicate ASA in machines. Some CASA systems are closely modelled on the known stages of auditory processing, whereas others adopt a more functional approach. However, all are broadly based on the principles underlying the perception and organisation of sound by human listeners, and in this respect they differ from ICA and other approaches to sound separation. In this paper, we review the principles underlying ASA and show how they can be implemented in CASA systems. We also consider the link between CASA and automatic speech recognition, and draw distinctions between the CASA and ICA approaches.

16.1 Introduction

Imagine a recording of a busy party, in which you can hear voices, music and other environmental sounds. How might a computational system process this recording in order to segregate the voice of a particular speaker from the other sources? Independent component analysis (ICA) offers one solution to this problem. However, it is not a solution that has much in common with that adopted by the best-performing sound separation system that we know of – the human auditory system. Perhaps the key to building a sound separator that rivals human performance is to model human perceptual processing?

This argument provides the motivation for the field of *computational auditory scene analysis (CASA)*, which aims to build sound separation systems that adhere to the known principles of human hearing. In this chapter, we review the state-of-the-art in CASA, and consider its similarities and differences with the ICA approach. We also consider the relationship between CASA and techniques for robust automatic speech recognition in noisy environments, and comment on the challenges facing this growing field of study.

Reprinted from *Speech Enhancement*, J. Benesty, S. Makino and J. Chen (Eds.), Springer, New York, 2005, pp. 371–402.

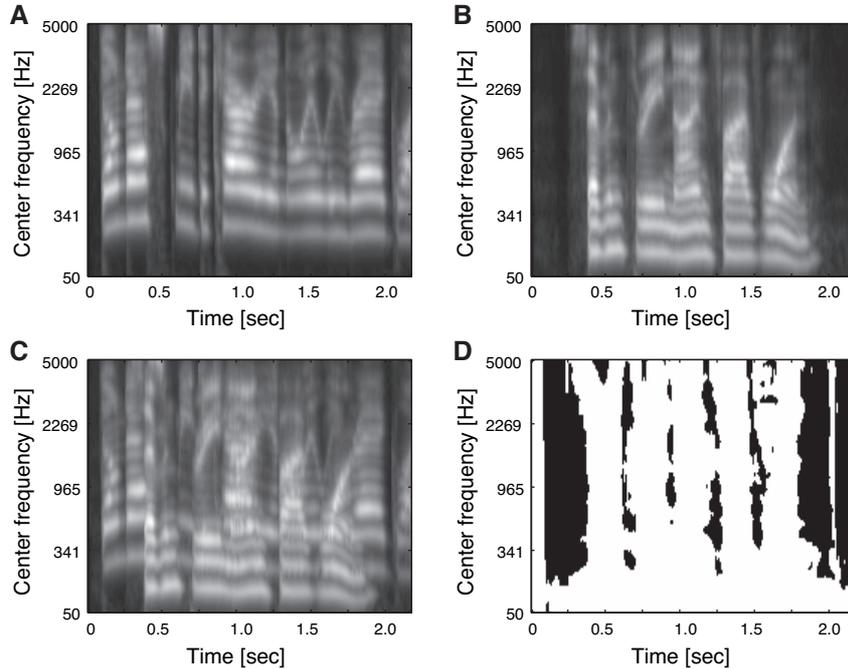


Fig. 16.1. (A) Auditory spectrogram for the utterance “don’t ask me to carry an oily rag” spoken by a female. (B) Auditory spectrogram for the utterance “seven three nine five one” spoken by a male. (C) Auditory spectrogram for a mixture of the male and female utterances. Light pixels correspond to regions of high energy, and dark pixels correspond to regions of low energy. (D) Ideal binary mask for the male utterance, obtained by using the criterion given in (16.1). White pixels indicate reliable regions, and black pixels indicate unreliable regions.

16.2 Auditory Scene Analysis

In naturalistic listening situations, several sound sources are usually active at the same time, and the pressure variations in air that they generate combine to form a mixture at the ears of the listener. A common example of this is the situation in which the voices of two talkers overlap, as illustrated in Figure 16.1C. The figure shows the simulated auditory nerve response to a mixture of a male and female voice, obtained from a computational model of auditory processing. How can this complex acoustic mixture be parsed in order to retrieve a description of one (or both) of the constituent sources?

Bregman [5] was the first to present a coherent answer to this question (see also [17] for a more recent review). He contends that listeners perform an *auditory scene analysis* (ASA), which can be conceptualised as a two-stage process. In the first stage, the acoustic mixture is decomposed into elements. An element may be regarded as an atomic part of the auditory scene, which

describes a significant acoustic event. Subsequently, a grouping process combines elements that are likely to have arisen from the same acoustic source, forming a perceptual structure called a *stream*. For example, consider the voice of a speaker; in Bregman's terms, the vocal tract of the speaker is the acoustic source, whereas the mental representation of the speaker's voice is the corresponding stream.

Grouping processes may be *data-driven* (primitive), or *schema-driven* (knowledge-based). In the former case, it is thought that listeners exploit heuristics similar to those proposed by the Gestalt psychologists for describing the ways in which elements of an image combine to form a coherent object. In schema-driven grouping, listeners apply learned knowledge of sound sources (such as speech and music) in a top-down manner. Examples of speech-related schemas include prosodic, semantic and pragmatic knowledge.

Consideration of Fig. 16.1C suggests that a number of primitive grouping cues could be applied to segregate the mixture. First, consider cues which might act upon acoustic components that overlap in time (so-called *simultaneous organisation*). At about 0.5 sec., the male speech begins and this gives rise to an abrupt onset in acoustic energy across all frequency channels. Hence, a principle of 'common onset' might allow the voices of the two speakers to be segregated in this region – frequency regions that exhibit an abrupt increase in energy at the same time are probably dominated by the same acoustic source. Similarly, a principle of 'common offset' could be applied to segregate the two speakers in the region close to 2 sec., when the male speech ceases. Another powerful grouping cue is harmonicity. In the figure, horizontal bands of energy are visible which correspond to harmonics of the same fundamental frequency (F0). In principle, these harmonics can be sorted into two sets, such that those related to the same F0 are grouped together.

Secondly, consider grouping cues which could act upon nonoverlapping acoustic components (so-called *sequential organisation*). In Fig. 16.1C, the male and female speakers occupy different average pitch ranges; hence the continuity of their F0s might be exploited in order to group successive utterances from the same speaker. Similarly, concentrations of energy in the time-frequency plane tend to change smoothly, such as those due to formant transitions. Again, such continuity can be exploited in order to separate one voice from the other. Some cues to sequential organisation are not illustrated by the figure. For example, listeners tend to group sounds that have a similar timbre, and which originate from the same location in space.

16.3 Computational Auditory Scene Analysis

The structure of a typical data-driven CASA system is closely related to Bregman's conceptual model, as shown in Fig. 16.2. In the first stage, the input mixture is processed in order to derive acoustic features. Subsequent grouping processes may operate directly on these features, or more usually

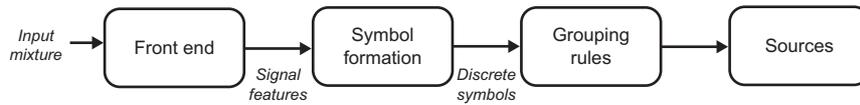


Fig. 16.2. Flow of processing in a typical data-driven CASA system, such as that of Brown and Cooke [7].

they will be used to derive an intermediate representation prior to grouping. In many systems, significant components in the time-frequency plane are encoded as discrete symbols. Grouping rules are then applied, in order to identify components that are likely to have arisen from the same source. The grouping heuristics may be encoded explicitly in a rule-based system, or may be implicitly encoded in a signal processing algorithm or neural network.

Once representations of individual sources are obtained, the auditory representation can usually be inverted in order to recover a time-domain waveform for the segregated source. This allows the separated signal to be evaluated using listening tests, or by performance metrics that involve a comparison between the original and reconstructed signals. Alternatively, an evaluation may be performed on the auditory representation directly.

An important notion in many CASA systems is the *time-frequency mask*. Given a description of the acoustic input in the time-frequency plane, a specific source may be recovered by applying a weighting to each time-frequency bin, such that regions dominated by the desired source receive a high weight and those dominated by other sources receive a low weight. The mask values may be binary or real-valued. Weintraub [67] was the first to use this approach in a CASA system, and it has since been adopted by several other workers [6,7,66,54,13]. The use of binary masks is motivated by the phenomenon of masking in human hearing, in which a weaker signal is masked by a stronger one within the same critical band (see Moore [41] for a review). It has also been noted that the reconstruction of a masked signal may be interpreted as a highly nonstationary Wiener filter [54].

What is the upper limit on the performance of a system that uses binary masks? Cooke *et al.* [13] have adapted a conventional speech recogniser so that reliable and unreliable (or missing) acoustic features are treated differently during decoding, and report excellent recognition performance using so-called *a priori* masks. Assuming that the clean speech and noise signals are available prior to mixing, the *a priori* mask is formed by selecting time-frequency regions in which the mixture energy lies within 3 dB of the energy in the clean speech. From the perspective of speech separation, Wang and colleagues [27,52,29] have subsequently proposed the *ideal binary mask* as a computational goal of CASA. Considering the auditory representation of a speech signal $s(t, f)$ and noise signal $n(t, f)$, where t and f index time and

frequency respectively, the ideal binary mask $m(t, f)$ is given by

$$m(t, f) = \begin{cases} 1 & \text{if } s(t, f) > n(t, f) \\ 0 & \text{otherwise} \end{cases} \quad (16.1)$$

A similar approach has been advocated by Jourjine *et al.* [31], who note that different speech utterances tend to be orthogonal in a high-resolution time-frequency representation, and can therefore be separated by binary masking. A number of workers have demonstrated that speech reconstructed from ideal binary masks is highly intelligible, even when extracted from a mixture of two or three concurrent speakers [54,52]. Speech intelligibility tests using both speech and babble noise interference also show that ideal binary masking can lead to substantial intelligibility improvements for human listeners [52]. An extensive discussion on ideal binary masks can be found in [65].

In the following sections, we first review the feature extraction stage of CASA, and then focus on monaural (one-microphone) and binaural (two-microphone) approaches. We also consider the issue of cue integration, and review a number of different computational frameworks that allow multiple grouping cues to be brought to bear on an acoustic signal.

16.3.1 Peripheral Auditory Processing and Feature Extraction

The first stage of a CASA system is usually a time-frequency analysis that mimics the frequency selectivity of the human ear. Typically, the input signal is passed through a bank of bandpass filters, each of which simulates the frequency response associated with a particular position on the basilar membrane. The ‘gammatone’ filter is often used, which is an approximation to the physiologically-recorded impulse responses of auditory nerve fibres [50,11]. The parameters of the gammatone filterbank (i.e., the filter order, bandwidth and frequency spacing) are usually chosen to provide a match to psychophysical data. Neuromechanical transduction in the cochlea may be approximated by half-wave rectifying and compressing the output of each filter; alternatively a detailed simulation of inner hair cell function can be employed [26]. We note, however, that not all CASA systems use an auditory-motivated time-frequency analysis. The short-term Fourier transform and discrete wavelet transform are also sometimes employed [42,56,38,43].

Examples of auditory spectrograms generated using a gammatone filterbank are shown in Fig. 16.1. Note that a nonlinear frequency scale is used, and that the bandwidth of each filter varies proportionately to its centre frequency. In low frequency regions, filter bandwidths are narrow and hence individual harmonics of a complex sound (such as speech) are resolved. In high-frequency regions, the bandwidths are broader and several components interact within the same filter.

Most CASA systems further process the peripheral time-frequency representation in order to extract features that are useful for grouping. The

motivation here is to explicitly encode properties which are implicit in the acoustic signal. Typical is the ‘synchrony strand’ representation of Cooke [11], which is a symbolic encoding of significant features in the time-frequency plane. Cooke demonstrates that the grouping stage of CASA (e.g., identifying harmonically related components) is facilitated by using a representation in which continuity in time-frequency is made explicit. A further example is the system of Brown [7], which forms representations of onset and offset events, periodicity and frequency transitions. Similar rich ‘mid level’ representations of the acoustic signal have been proposed by other workers [21,66].

16.3.2 Monaural Approaches

Although binaural cues contribute substantially to ASA, human listeners are able to segregate sounds when listening with a single ear, or when listening diotically to a single-channel recording. Perceptually, one of the most potent cues for monaural sound segregation is fundamental frequency (F0); specifically, listeners are able to exploit a difference in F0 in order to segregate the harmonics of one sound from those of interfering sounds. Accordingly, much of the work on monaural CASA has focussed on the problem of identifying the multiple F0s present in an acoustic mixture (so-called ‘multipitch analysis’), and using them to separate the constituent sounds. Perhaps the earliest example is the system for separating two concurrent speakers described by Parsons [49]. In his approach, the harmonics of a target voice are selected by peak picking in the spectral domain, and the voice of each speaker is tracked using pitch continuity.

An important class of algorithms for F0 estimation is based on a temporal model of pitch perception proposed by Licklider [36]. The first computational implementation of Licklider’s theory was described by Weintraub [67], who referred to it as an ‘auto-coincidence’ representation; subsequently, Slaney and Lyon [57] introduced the term *correlogram*. The correlogram is computed in the time domain by performing an autocorrelation at the output of each channel of a cochlear filter analysis,

$$A(t, f, \tau) = \sum_{n=0}^{N-1} h(t-n, f)h(t-n-\tau, f)w(n) \quad (16.2)$$

Here, $h(t, f)$ represents the cochlear filter response for channel f at time frame t , τ is the autocorrelation delay (lag), and w is a window function of length N samples (typically a Hanning, exponential or rectangular window is used). Alternatively, the autocorrelation may be performed in the frequency domain by means of the discrete Fourier transform (DFT) and its inverse transform (IDFT), i.e.

$$\text{IDFT}(|\text{DFT}(h)|^k) \quad (16.3)$$

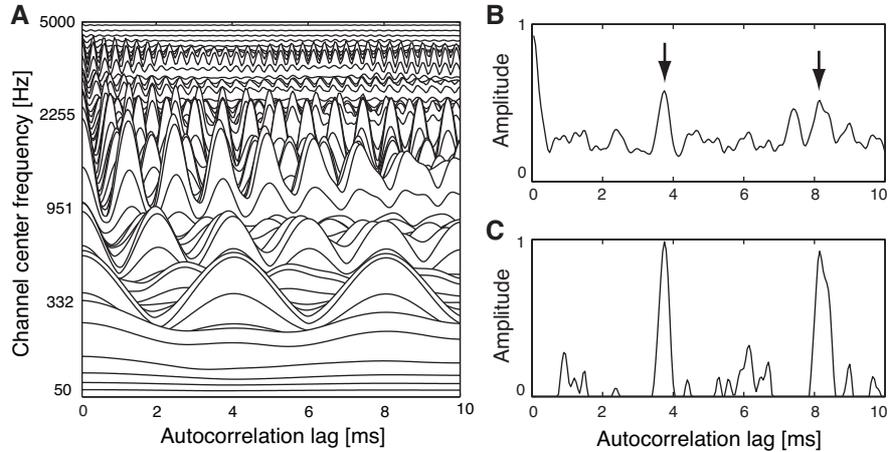


Fig. 16.3. A. Correlogram for time frame 100 of the mixture of two speakers shown in Fig. 16.1. B. Summary autocorrelation function (SACF). The pitch periods of the two speakers are marked with arrows. The male voice has a period of 8.1 ms (corresponding to a F0 of 123 Hz) and the female voice has a period of 3.8 ms (corresponding to a F0 of 263 Hz). C. Enhanced SACF, in which one iteration of processing has been used to remove sub-octave multiples of the significant peaks.

where h is a windowed section of the cochlear filter response. The introduction of a parameter k allows for a ‘generalised autocorrelation’ [62]. For conventional autocorrelation $k = 2$, but smaller values of k are advantageous because this leads to sharper peaks in the resulting function ([62] suggest a value of $k = 0.67$).

The correlogram is an effective means for F0 estimation because it detects the periodicities present in the output of the cochlear filterbank. For example, consider a voice with a F0 of 125 Hz. A channel responding to the fundamental component of the voice has a period of 8 ms, and hence a peak occurs in the corresponding autocorrelation function at a lag of 8 ms. Similarly, a channel responding to the second harmonic (250 Hz) has an autocorrelation peak at 4 ms, but because of the periodic nature of the autocorrelation function, peaks also occur at 8 ms, 12 ms, 16 ms and so on. In high-frequency regions, cochlear filters are wider and a number of harmonics interact within the same filter, causing amplitude modulation (AM). These interacting components ‘beat’ at a rate corresponding to the fundamental period, and also cause a peak in the autocorrelation function at the corresponding lag. Hence, for a periodic sound a ‘spine’ occurs in the correlogram which is centered at the fundamental period (8 ms in our example); for an example, see Fig. 16.3A. A convenient means of emphasizing this F0-related structure is to sum the channels of the

correlogram over frequency,

$$S(t, \tau) = \sum_{f=1}^M A(t, f, \tau) \quad (16.4)$$

The resulting *summary autocorrelation function (SACF)* $S(t, \tau)$ exhibits a peak at the period of each F0, and can be used as the basis for multipitch analysis (Fig. 16.3B). For example, Tolonen and Karjalainen [62] describe a computationally efficient multipitch model based on the SACF. Computational savings are made by splitting the input signal into two bands (below and above 1 kHz) rather than performing a multi-band frequency analysis. A generalized autocorrelation is then computed for the low-frequency band and for the envelope of the high frequency band, and added to give a SACF. Further processing is then performed to enhance the representation of different F0s. Specifically, the SACF is half-wave rectified and then expanded in time by a factor of two, subtracted from the original SACF and half-wave rectified again. This removes peaks that occur at sub-octave multiples, and also removes the high-amplitude portion of the SACF close to zero delay (Fig. 16.3C). The operation may be repeated for time expansions of a factor of 3, 4, 5 and so on, in order to remove higher-order multiples of significant pitch peaks. In [32], the authors show how pitch tracks from this system can be used to separate harmonic sounds (two vowels) by applying a comb-notch filter, which removes the harmonics of the pitch track to which it is tuned. Ottaviani and Rocchesso [46] also describe a speech separation system based on F0 tracking using the enhanced SACF. They resynthesize a separated speech signal from a highly zero-padded Fourier spectrum, which is selectively weighted to emphasize harmonics of the detected pitch.

One of the more sophisticated algorithms for tracking the pitch of multiple speakers is reported by Wu *et al.* [69]. Their approach consists of four stages, shown schematically in Fig. 16.4. In the first stage, the digitised input signal is filtered by a bank of gammatone filters, in order to simulate cochlear filtering. Further stages of processing treat low-frequency channels (which have a centre frequency below 800 Hz) and high-frequency channels differently. In low-frequency channels the correlogram is computed directly from the filter outputs, whereas in high-frequency channels the envelope in each channel is autocorrelated.

In the second stage of the system, ‘clean’ correlogram channels (i.e., those that are likely to contain reliable information about the periodicity of a single sound source, and are relatively uncorrupted by noise) are identified. The third stage of the system estimates the pitch periods present in each individual time frame using a statistical approach. Specifically, the difference between the true pitch period and the time lag of the closest correlogram peaks in each channel is employed as a means of quantifying the support for a particular pitch period hypothesis.

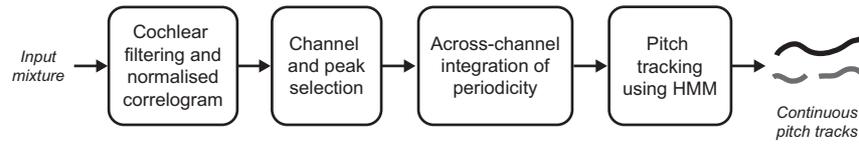


Fig. 16.4. Schematic diagram of the Wu, Wang and Brown [69] system for tracking multiple fundamental frequencies in an acoustic mixture.

Periodicity information is then integrated across channels in order to derive the conditional probability of observing a set of pitch peaks $P(\Phi|x)$ given a pitch state x . Since zero, one or two pitches may be present, the pitch state is regarded as a pair $x = (y, Y)$ where $y \in R^Y$ is the pitch period and $Y \in \{0, 1, 2\}$ is the space index. Channel conditional probabilities are combined into frame conditional probabilities by assuming the mutual independence of the responses in all channels.

In the final stage of Wu *et al.*'s system, pitch periods are tracked across time using a hidden Markov model (HMM). Hidden nodes in the HMM represent the possible pitch states in each time frame, and observation nodes represent the set of selected peaks in each time frame. Transition probabilities between time frames are estimated from a small corpus of speech signals. Transition probabilities between state spaces of zero, one and two pitches are also estimated from the same corpus of speech signals, with the assumption that a single speaker is present for half of the time and two speakers are present for the remaining time. The optimal state sequence is found by the Viterbi algorithm, and may consist of zero, one or two pitch states. Fig. 16.5 shows an example of the F0 tracks derived by Wu *et al.*'s system for a mixture of two speakers.

Wu *et al.*'s system suffers from a number of limitations. In principle, the algorithm could be modified to track more than two simultaneous speakers by considering more than three pitch spaces, but it remains to be seen how well such an approach would work in practice. Although it is robust to the presence of an interfering speaker (and can track its F0), Khurshid and Denham [35] find that Wu *et al.*'s system is less robust in the presence of background noise. They also find that although Wu's algorithm tracks the F0 of the dominant speaker in a mixture very accurately, its estimate of the nondominant F0 can be poor.

Khurshid and Denham suggest an alternative approach that rectifies some of these problems, which is based on the analysis of the output of a bank of damped harmonic oscillators, which model the frequency analysis performed by the cochlea. Analysis of the fine time structure (consecutive zero crossings and amplitude peaks) of each oscillator output is performed in order to determine the driving frequency. An algorithm is then used to hypothesise the F0 (or multiple F0s) that are present in order to explain the observed frequency components. This is achieved by identifying salient spectral peaks

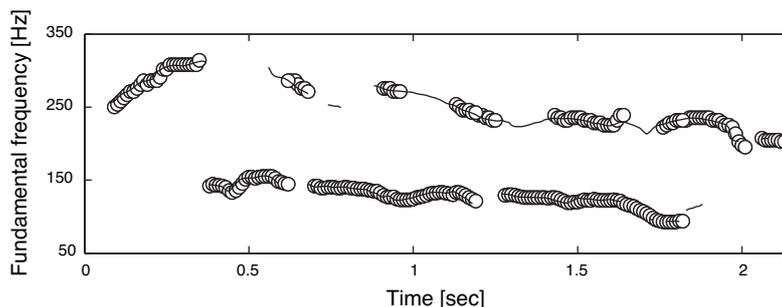


Fig. 16.5. Pitch tracks for the mixture of two speech signals shown in Fig. 16.1, obtained using the algorithm of Wu, Wang and Brown [69]. Solid lines show the ground-truth pitch tracks for each speaker, open circles show the estimated pitch periods at each time frame.

(similar to the place groups described by Cooke [13]) and then assessing the support for every subharmonic of the peak that falls within the normal range of voice pitch. Such a frequency ‘remapping’ leads to noise robustness, and may be regarded as a simple model of simultaneous masking in the auditory nerve. Simple continuity constraints are used to track two F0s over time. Khurshid and Denham performed a comparison and reported that, although Wu *et al.*’s system is able to more accurately track the dominant pitch, their own system tracks the nondominant F0 more reliably and is also more robust to noise.

The correlogram, as described in (16.2), is based on an autocorrelation operation such that a large response occurs at the period of a F0. An alternative approach, advocated by de Cheveigné [15], is to perform *cancellation* rather than autocorrelation. In his approach, a time-domain comb filter of the form

$$q(t) = \delta(t) - \delta(t - \tau) \quad (16.5)$$

is applied to the acoustic signal (or to the output from each channel of a cochlear filter bank), where $\delta(t)$ is the delta function, t is time and τ is the lag parameter. The filter has zeros at frequencies $f = 1/\tau$ and all its multiples, and hence its response to a signal with a period of τ is zero. F0 analysis is therefore performed by applying the filter for different values of τ and searching for the minimum response. de Cheveigné and Kawahara [16] further suggest that this approach can be extended to the problem of multipitch estimation by cascading N filters, each of which is tuned to cancel a particular period. Hence, to perform multipitch estimation it is simply necessary to search the N -dimensional space of lag parameters until the minimum is found. The authors evaluated their algorithm on a corpus consisting of mixtures of two or three harmonic complexes, with impressive results. However, their joint cancellation technique has certain limitations. It is computation-

ally expensive (although amenable to parallelism), and cancellation of one source may partially cancel another source if their periods are related by an integer multiple.

As noted previously, the correlogram deals with two cues to the F0 of a sound in a unified way; resolved harmonics in low-frequency regions and AM (‘beating’) in high-frequency regions. However, this unified treatment leads to poor segregation in the high-frequency range because AM alters autocorrelation structure and makes it difficult to group high-frequency components [29]. Some CASA systems process resolved and unresolved harmonic regions using different mechanisms (e.g., [11]). Hu and Wang [29] describe a recent system in which AM is extracted in high frequency regions and used to segregate unresolved harmonics, whereas conventional techniques are used to segregate resolved harmonics. AM detection is based on an ‘envelope correlogram’, which is of the form given in (16.2) except that the autocorrelation is performed on the envelope of each filter response rather than the fine structure. Correlations between adjacent channels are then computed in order to identify significant acoustic components. An initial segmentation based on F0 is then performed, which is similar to that described by [66]. Further processing is used to refine the F0 track for the dominant source, based on temporal smoothness and a periodicity constraint. Time-frequency units are then labelled according to whether they are dominated by the target speech signal or not, using heuristics that are based on the conventional correlogram in low-frequency regions and the envelope correlogram in high frequency regions. Finally, segments are generated based on cross-channel envelope correlation and temporal continuity, and these are grouped with low-frequency segments that share a common F0. The authors show that their system performs consistently better than that of Wang and Brown [66] across 10 noise conditions. In all but one noise condition it also outperforms a conventional spectral subtraction scheme for speech enhancement.

Explicit representations of AM have been proposed as an alternative to the correlogram. For instance, Berthommier and Meyer [2] describe a system for separating sounds on the basis of their F0s using the modulation spectrum (see also [34]). Each channel of a gammatone filterbank is half-wave rectified and bandpass filtered to remove the DC component and frequencies above the pitch range. The magnitude of a DFT is then computed to give a two-dimensional representation of tonotopic frequency against AM frequency. A harmonic sieve is then applied to perform F0 analysis and grouping according to common F0. In a subsequent paper [3], the authors extend their system by introducing two further stages of processing. The first of these addresses a problem caused by the distributive nature of the DFT, namely that evidence for a particular F0 is distributed across various harmonic frequencies along the modulation frequency axis of the map. The author’s solution is to compute a pooled map, in which evidence for each F0 is integrated. The resulting representation is better suited to grouping and pitch analysis, since

a single peak occurs in the pooled map for each period source. The second stage of processing is an identification map, which estimates the correlation between stored spectral prototypes and each spectral slice along the modulation frequency axis. This allows classification of vowel spectra without the need for an explicit F0 detection stage. It is an open question whether a similar mechanism could be used to segregate continuous speech, rather than isolated vowels; the computational cost may be prohibitive.

Other principles of auditory organization, such as spectral smoothness, may also be used to improve F0 detection and tracking. Klapuri [33] describes a multipitch estimation technique which exploits a spectral smoothness principle. His system uses an iterative approach to multipitch estimation, in which a predominant F0 is found, and then the corresponding harmonic spectrum is estimated and linearly subtracted from the mixture. This process is then repeated for the residual. However, this approach has a tendency to make errors when constituent F0s in the mixture are harmonically related, because cancellation of one F0 may inadvertently remove a frequency component that is shared by another source. The solution proposed in [33] is to smooth the spectrum before subtraction; partials containing energy from more than one source extend above the smoothed envelope, so that they are preserved in the residual when the smoothed envelope is subtracted from the mixture. Klapuri shows that application of the spectral smoothness constraint reduces the error rate for pitch analysis of four-pitch mixtures by about half.

Finally, although most monaural CASA systems have used F0-based cues, there have been some attempts to exploit other cues such as frequency modulation [40] and common onset [18], [6], [7], [28]. For example, Denbigh and Zhao [18] describe a system which selects the harmonics for a target voice in a manner that is similar to the approach described by Parsons [49]. Additionally, their system compares adjacent spectra in order to determine whether the onset of a new voice has occurred. This allows their pitch tracking algorithm to extract weak voiced sounds, and increases the accuracy of pitch tracking when two voices are active. Common onset is currently a somewhat under-utilised cue in CASA systems, although a recent study has been described in which it is employed to segregate stop consonants [28].

16.3.3 Binaural Approaches

The principal cues that human listeners use to determine the location of a sound source are those that involve a comparison between the two ears. A sound source located to one side of the head generates sound pressure waves that arrive at the nearer ear slightly before the farther ear; hence there is an interaural time difference (ITD) which provides a cue to source location. Similarly, the sound intensity will be greater in the nearer ear, causing an interaural intensity difference (IID). The IID is usually expressed in decibels, in which case it is termed the interaural level difference (ILD). The relative efficacy of ITD and ILD cues depends on frequency. At low frequencies,

sounds diffract around the head and hence there is no appreciable ILD below about 500 Hz. At high frequencies, ITD does not provide a reliable cue for the location of tonal sounds because of phase ambiguities. However, the envelope of complex sounds can be compared at the two ears in high frequency regions; this cue is referred to as the interaural envelope difference (IED).

In addition to binaural comparisons, direction-dependent filtering by the head, torso and pinnae provide cues to source location. These provide some ability to localise sounds monaurally, and are particularly important for discrimination of elevation and for resolving front-back confusions. Such cues are seldom used explicitly by CASA systems and are not considered here; however, their use in CASA systems remains an interesting area for future research. Preliminary work on a sound localization system that exploits pinna cues is reported in [25].

Computational systems for binaural signal separation have been strongly influenced by two key ideas in the psychophysical literature. The first is Durlach's [20] equalization-cancellation (EC) model of binaural noise suppression, which is a two-stage scheme. In the first stage, equalization is applied to make the noise components identical in each of two binaural channels. This is followed by a cancellation stage, in which the noise is removed by subtracting one channel from the other. Many two-microphone approaches to noise cancellation may be regarded as variants of the EC scheme (e.g., [61], [38], [56]).

The second key idea motivating binaural signal separation systems is the cross-correlation model of ITD processing proposed by Jeffress [30]. In this scheme, neural firing patterns arising from the same critical band of each ear travel along a dual delay-line system, and coincide at a delay corresponding to the ITD. Computationally, the Jeffress model may be expressed as a cross-correlation of the form

$$C(t, f, \tau) = \sum_{n=0}^{N-1} h_L(t-n, f)h_R(t-n-\tau, f)w(n) \quad (16.6)$$

where $h_L(t, f)$ and $h_R(t, f)$ represent the simulated auditory nerve response in the left and right ears respectively for time frame t and frequency channel f , and $w(n)$ is a window of size N samples. The resulting cross-correlogram $C(t, f, \tau)$ is closely related to the correlogram given in (16.2); both are three-dimensional representations in which frequency, time and lag are represented on orthogonal axes. Fig. 16.6A shows a cross-correlogram for a mixture of a male and female speaker, originating from azimuths of -15 degrees and +10 degrees respectively. As with the correlogram, it is convenient to sum the cross-correlation functions in each frequency band to give a summary cross-correlation function (SCCF), in which large peaks occur at the ITD of each source (Fig. 16.6B). The figure also shows the ILD for this mixture,

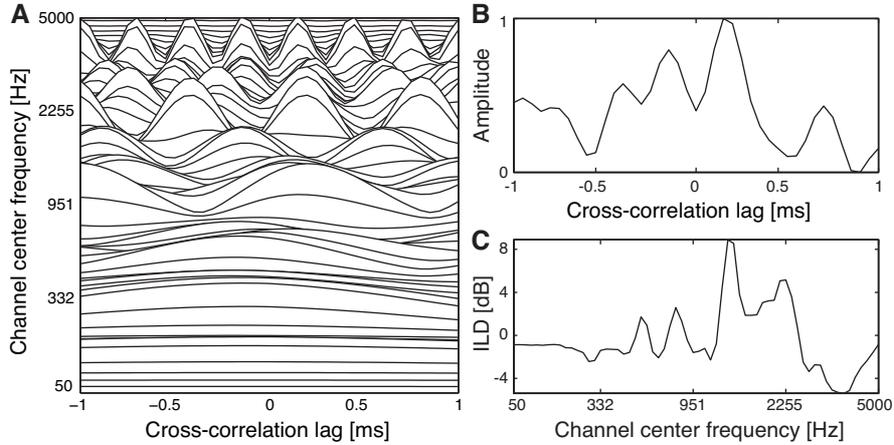


Fig. 16.6. A. Cross-correlogram for time frame 100 of the mixture of speakers shown in Fig. 1, for which the male speaker has been spatialised at an azimuth of -15 degrees and the female speaker at an azimuth of +10 degrees. B. Summary cross-correlogram. C. Interaural level difference (ILD) in each frequency channel.

computed using

$$ILD(t, f) = 10 \log_{10} \left(\frac{\sum_{n=0}^{N-1} h_R(t+n, f)^2}{\sum_{n=0}^{N-1} h_L(t+n, f)^2} \right) \quad (16.7)$$

Note that, as expected, the ILD is negligible in low frequency channels. However, in the mid-frequency region channels tend to be dominated by the female speaker and exhibit a large positive ILD. Above 2.5 kHz, the male speaker is dominant and a substantial negative ILD is observed.

Early attempts to exploit spatial cues in a system for speech segregation include the work of Lyon [39] and the binaural ‘cocktail party processor’ described by Bodden [4]. Bodden’s system is based on a cross-correlation mechanism for localising the target and interfering sources, and uses a time-variant Wiener filter to enhance the target source. Effectively this filter applies a window function to the azimuth axis of the cross-correlogram, such that energy from a speaker at a target azimuth is retained, and the remainder is cancelled. Bodden reports good performance for mixtures of two or three speakers in anechoic conditions.

Bodden’s system uses a modification of the Jeffress scheme in which contralateral inhibition is employed to sharpen the cross-correlation pattern. Numerous other approaches have been described for improving the accuracy of location estimates from cross-correlation processing, such as the ‘stencil’ filter proposed by Liu *et al.* [37]. The SCCF shown in Fig. 16.6B is a direct way of estimating the ITD of a sound source from the cross-correlogram, but it assumes that the ITD is independent of frequency. This assumption does

not hold if the binaural recordings are obtained from a dummy head, because diffraction around the head introduces a weak frequency dependence to the ITD. The ‘stencil’ approach described by Liu *et al.* is a more sophisticated way of determining source location, which is based on pattern-matching the peaks in the cross-correlogram. At the ITD of a sound source, the pattern of peaks in the cross-correlogram exhibits a structure in which curved traces fan out from a central vertical line (two such structures are visible in Fig. 16.6A, which is a cross-correlogram for a two-source mixture). Accordingly, Liu *et al.* derive a SCCF by integrating activity in the cross-correlogram over a template (‘stencil’) for each location, which matches the expected pattern of peaks. They report enhanced localization of sources in the azimuthal plane using this method.

A related approach is described by Palomäki *et al.* [47], who form a ‘skeleton’ cross-correlogram by identifying local peaks and replacing each with a narrower Gaussian. This avoids the problem of very wide peaks, which occur in low-frequency channels and bias the location estimates in the SCCF (see Fig. 16.6A). Furthermore, in the process of forming the skeleton cross-correlogram, peak positions are mapped from ITD to an azimuth axis using frequency-dependent look-up tables. Again, this overcomes the problems associated with the frequency dependence of ITD. Palomäki *et al.*’s system also includes a mechanism for reducing the effect of echoes on localization estimates. This consists of a delayed inhibition circuit, which ensures that location cues at the onset of a sound source have more weight than those that arrive later. In this respect, it may be regarded as a simple model of the precedence effect (for a review, see [41]). The authors report that use of the inhibition mechanism improves the robustness of source localization in mildly reverberant environments.

Roman *et al.* [52] describe a binaural speech separation algorithm which is based on location estimates from skeleton cross-correlograms. They observe that, within a narrow frequency band, modifying the relative strength of a target and interfering source leads to systematic changes in the observed ITD and ILD. For a given location, the deviation of the observed ITD and ILD from ideal values can therefore be used to determine the relative strength of the target and interferer, and in turn this can be used to estimate the ideal binary mask (see (16.1)). Specifically, a supervised learning method is used for different spatial configurations and frequency bands based on an ITD-ILD feature space. Given an observation x in the ITD-ILD feature space, two hypotheses are tested for each channel; whether the target is dominant (H_1) and whether the interferer is dominant (H_2). Based on estimates of the bivariate densities $p(x|H_1)$ and $p(x|H_2)$, classification is performed using a maximum *a posteriori* (MAP) decision rule, i.e. $p(H_1)p(x|H_1) > p(H_2)p(x|H_2)$. Roman *et al.*’s system includes a resynthesis pathway, in which the target speech signal is reconstructed only from those time-frequency regions selected in the binary mask. They report a performance in anechoic environments which is

very close to that obtained using the ideal binary mask, as determined using three different evaluation criteria (signal-to-noise ratio (SNR), automatic speech recognition accuracy and listening tests).

A limitation of binaural systems is that they generally perform well when two sources are present, but their performance degrades in the presence of multiple interferers. Liu *et al.* [38] describe a multi-band mechanism which allows this limitation to be overcome to some extent. Their binaural cancellation scheme is based on a subtraction of the two input signals. Essentially, their system generates a nulling pattern for each point on the lag-axis of the cross-correlogram, such that the null occurs at the direction of the noise source and unity gain is maintained in the direction of the target sound. An innovative aspect of their approach is that the null in each frequency band can be steered independently, so that at each time instant it cancels the noise source that emits the most energy in that band. This allows their system to cancel multiple noise sources, provided that their locations are known; this information is provided by the author's system for sound localization, discussed above. Liu *et al.* show that when four talkers are present in an anechoic environment, their system is able to cancel each of the three interfering speakers by 3-11 dB whilst causing little degradation to the target speaker. Similar results were obtained for a six-talker scenario. However, in a moderately reverberant room the total noise cancellation fell by about 2 dB; this raises doubts as to whether the system is sufficiently robust for use in real-world acoustic environments.

A number of workers have developed systems that combine binaural processing with other grouping cues (usually those related to periodicity). An early example is the system proposed by Kollmeier and Koch [34]. They describe a speech enhancement algorithm which works in the domain of the modulation spectrum, i.e. a two-dimensional representation of AM frequency vs. center frequency. Energy from each sound source tends to form a cluster in the modulation spectrum, allowing sources with different modulation characteristics to be separated from one another. Binaural information (ITD and ILD cues) is used to suppress clusters that do not arise from a desired spatial location.

Related speech separation systems which use F0 and binaural cues are described by Denbigh and colleagues [18], [56]. In the latter approach, the cross-correlation between two microphones is continuously monitored in order to determine the azimuth of the most intense sound in each time frame. If the most intense sound lies close to the median plane, it is assumed to be speech and an initial estimate is made of the speech spectrum. An estimate of the total interference (noise and reverberation) is also obtained by cancelling the dominant target signal. Subsequent processing stages refine the estimate of the speech spectrum, by subtracting energy from it that is likely to be contributed by the interference. Finally, F0 analysis is performed on the extracted target signal, and cross-referenced against F0 tracks from the left and

right microphones. A continuity constraint is then applied to ensure that the F0 of the estimated target speech varies smoothly. The target speech signal is reconstructed using the overlap-add technique and passed to an automatic speech recogniser. The authors report a large gain in ASR accuracy for an isolated word recognition task in the presence of a speech masker and mild reverberation; accuracy increased from 30% to 95% after processing by the system, for a SNR of 12 dB.

Okuno *et al.* [45] also describe a system which combines binaural and F0 cues, and they assess its ability to segregate mixtures of two spatially separated speakers. Harmonic fragments are found in each of the left and right input channels, and then a direction is computed for pairs of fragments using ITD and ILD cues. The authors evaluate their system on a speech recognition task, but focus on the ability of the system to recognize *both* utterances rather than a single target utterance. They find that ASR error rates are substantially reduced by using their system, compared to performance on the unprocessed speech mixtures. They also report that binaural cues play an important role in this result; the ASR accuracy of a system which only used harmonic fragments was about half that of a system which used both harmonic and binaural cues.

16.3.4 Frameworks for cue integration

So far, we have focused on the cues that are pertinent to CASA, but a key issue remains – how can cues be combined in order to find organisation within an acoustic signal, and hence retrieve a description of a target sound source from a mixture?

The earliest approaches to CASA were motivated by classical artificial intelligence techniques, in that they emphasised representation and search. For example, Cooke's system [11] employs a synchrony strand representation, in which the acoustic scene is encoded as a collection of symbols that extend through time and frequency. A search algorithm is employed to identify groups of strands that are likely to have arisen from the same source. This is mainly achieved on the basis of harmonicity; search proceeds from a 'seed' strand, and other strands that are harmonically related to the seed strand are added to its group. In a second grouping stage a pitch contour is derived for each group, using the frequency of resolved harmonics in low-frequency regions, and using AM frequency in high-frequency regions. Groups that share a common pitch contour are then combined. In addition, a subsumption stage removes groups whose strands are contained in a larger grouping. Brown [6,7] describes a similar approach, but substantially expands the palette of acoustic representations by including time-frequency 'maps' of onset activity, offset activity, frequency transition and periodicity. These are combined to form a symbolic representation of the acoustic scene, which is searched in a similar manner to Cooke's.

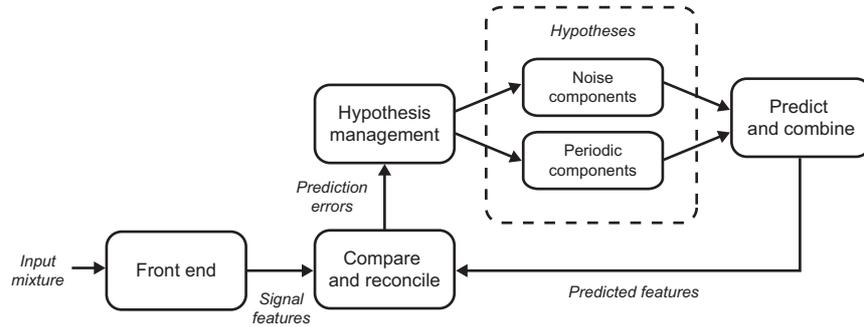


Fig. 16.7. Flow of processing in the prediction-driven architecture of Ellis [21]. Redrawn from [23].

Neither of the systems described above constitute a generic architecture for cue integration. Rather, Cooke's system combines groups of strands using a single derived property (pitch contour) and Brown's system performs cue integration during the formation of time-frequency objects. Hence, it is not clear how other cues (such as those relating to spatial location) could be included in these systems. Also, both are essentially data-driven architectures, as shown in Fig. 16.2. In general, solution of the CASA problem requires the application of top-down knowledge as well as bottom-up processing. Finally, both systems run in 'batch' mode; they process the acoustic signal in its entirety in order to derive an intermediate representation, which is then searched. Clearly, an architecture that allows real-time processing is necessary for most applications (such as hearing prostheses and automatic speech recognition).

A more generic architecture for cue integration is the blackboard, as advocated by Cooke *et al.* [12] and Godsmark and Brown [24]. In this scheme, grouping principles such as harmonicity are cast as knowledge sources ('experts') that communicate through a globally accessible data structure (the blackboard). Experts indicate when they are able to perform an action, and place their results back on the blackboard. For example, a harmonicity expert might be initiated because harmonically related partials are available on the blackboard, and would compute a pitch contour from them. In turn, another expert might combine groups that have the same pitch contour. Centralised control is provided by a scheduler, which determines the order in which experts perform their actions. Blackboard architectures are well suited to CASA because they were developed to deal with problems that have a large solution space, involve noisy and unreliable data, and require many semi-independent sources of knowledge to form a solution.

Godsmark and Brown's system [24] is specialised for musical signals, rather than speech, but is interesting because it suggests a mechanism for resolving competition between grouping principles. Such competition might

arise if, for example, two acoustic components were sufficiently distant in frequency to be regarded as separate, but sufficiently close in time to be regarded as grouped. They adopt a ‘wait and see’ approach to this problem; many possible organisations are maintained within a sliding time window. Within the sliding window, alternate organisations of synchrony strands are scored by grouping experts. An organisation is only imposed on a section of the acoustic signal after the window has passed over it, thus allowing contextual information to influence the organisation of strands into groups. The authors also show how top-down and bottom-up processing can be combined in a multi-layered blackboard architecture. For example, predictions about anticipated events, based on a previously observed temporal pattern, can be used to influence the formation and grouping of synchrony strands.

A similar emphasis on the role of top-down processing is found in the study by Ellis [21], who describes a *prediction-driven* architecture for CASA. By way of contrast with the Cooke and Brown systems, in which the flow of information is linear and data-driven, Ellis’s approach involves a feedback loop so that predictions derived from a ‘world model’ can be compared against the input (see Fig. 16.7). The front-end processing of Ellis’ system forms two representations, a time-frequency energy envelope and correlogram. These representations are reconciled with predictions based on world-model hypotheses by a comparison block. The world model itself consists of a hierarchy of increasingly specific sound source descriptions, the lowest level of which is couched in terms of three sound elements; wefts (which represent pitched sounds), noise clouds and transient clicks. A reconciliation engine, which is based on a blackboard system, updates the world model according to differences detected between the observed and predicted signals.

Okuno *et al.* [44] describe a residue-driven architecture for CASA which is closely related to Ellis’ approach, in that it compares the acoustic input against predictions from a world model. However, Ellis’ system makes this comparison at the level of intermediate acoustic representations (such as the smoothed spectral envelope). In contrast, the residue-driven architecture reconstructs a time-domain waveform for the modelled signal components, and subtracts this from the acoustic input to leave a residue which is then further analysed. Okuno *et al.* implement the residue-driven approach within a multi-agent system, in which three kinds of agent (event-detectors, tracer-generators and tracers) initiate and track harmonic fragments. The multi-agent framework is similar in concept to the blackboard – ‘experts’ and ‘agents’ are roughly equivalent – except that agents communicate directly rather than through a global data structure.

Some recent approaches to cue integration in CASA have been motivated by the development of powerful algorithms in the machine learning community, rather than classical artificial intelligence techniques. For example, Nix *et al.* [43] describe a statistical approach to CASA which is based on a state-space approach. Specifically, they consider the problem of separating three

speakers using two microphones. The problem is formulated as a Markov state-space of the form

$$\mathbf{x}_k = \mathbf{f}_k(\mathbf{x}_{k-1}, \mathbf{v}_{k-1}) \quad (16.8)$$

$$\mathbf{z}_k = \mathbf{g}_k(\mathbf{x}_k, \mathbf{n}_k) \quad (16.9)$$

Here, \mathbf{x}_k represents the azimuth, elevation and short-time magnitude spectrum of each speaker at time k (which are unknown), and \mathbf{z}_k is the power spectral density observed at the two microphones. The function $\mathbf{f}_k(\mathbf{x}_{k-1}, \mathbf{v}_{k-1})$ corresponds to the probability density function $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, i.e. the probability that one pair of directions and magnitude spectra at time k succeed another pair of directions and magnitude spectra at time $k-1$. The function $\mathbf{g}_k(\mathbf{x}_k, \mathbf{n}_k)$ corresponds to $p(\mathbf{z}_k|\mathbf{x}_k)$, which is the probability that an observation \mathbf{z}_k is made when the state of the system is \mathbf{x}_k . The random variables \mathbf{v}_k and \mathbf{n}_k are termed the ‘process noise’ and ‘observation noise’ respectively, and have known statistics. The task is to estimate \mathbf{x}_k from the values of \mathbf{z}_i , $1 \leq i \leq k$, in an optimal manner. This estimation task is performed by a sequential Monte Carlo method (also known as the ‘particle filter’ or ‘condensation’ algorithm).

The performance of the system reported in [43] is somewhat disappointing; although the system reliably tracks the direction and short-time magnitude spectrum of a single source, it is unable to estimate the spectra of two concurrent voices with any accuracy. Additionally, the computational requirement of the algorithm is high and training is time consuming; the authors report that it took several weeks to estimate $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ from a large database of recorded speech.

Finally, we note that the problem of cue integration in CASA is closely related to the *binding problem*. This term refers to the fact that information about a single sensory event is distributed across many areas of the brain – how is this information bound together to form a coherent whole? One possibility is that the grouping of neural responses is performed by *oscillatory correlation*. In this scheme, neurons that represent features of the same sensory event have synchronised responses, and are desynchronised from neurons that represent different events. Wang and colleagues [64,66] have used the principle of oscillatory correlation to build a neurobiologically-motivated architecture for CASA. In the first stage of their scheme, the correlogram is employed to detect the periodicities present in local time-frequency regions. Subsequently, processing is performed by a two-layer oscillator network which mirrors the two conceptual stages of ASA. In the first (segmentation) layer, periodicity information is used to derive segments, each of which encodes a significant feature in the time-frequency plane. Oscillators belonging to the same segment are synchronised by local connections. In the second (grouping) layer, links are formed between segments that have compatible periodicity information along their length (i.e., those that are likely to belong to the same F0). As a result, groups of segments form in the second layer which correspond to sources that have been separated by their F0.

Frameworks for CASA based on neural oscillators have two attractive features. Firstly, they are based on a parallel and distributed architecture which is suitable for implementation in hardware. Secondly, because the oscillator in each time-frequency region may be regarded as ‘on’ or ‘off’ at any particular time instant, the output of an oscillator array may be interpreted as a binary time-frequency mask. This makes them eminently suitable as a front-end to ASR systems that employ missing feature techniques (see below).

16.4 Integrating CASA with Speech Recognition

Conventional ASR systems are constructed on the assumption that the input to them will be speech. In practice this is usually not the case, because speech is uttered in acoustic environments in which other sound sources may be present. As a result, the performance of conventional ASR system declines sharply in the presence of noise.

ASR is a pattern recognition problem in which observed acoustic features \mathbf{X} must be assigned to some class of speech sound. This is achieved by selecting the word sequence W which maximises the posterior probability $P(W|\mathbf{X})$, which can be expressed using Bayes theorem as

$$\hat{W} = \operatorname{argmax}_w \frac{P(\mathbf{X}|W)P(W)}{P(\mathbf{X})} \quad (16.10)$$

where $P(W)$ is the language model and the likelihood $P(\mathbf{X}|W)$ is the acoustic model. One approach to improving the noise-robustness of ASR is to enhance the speech in the acoustic mixture, so that the observed features resemble the acoustic model as closely as possible. This provides the most straightforward approach to integrating CASA and ASR; the CASA system segregates the speech from the acoustic mixture, and then a ‘clean’ signal is resynthesized and passed to the recogniser. A number of CASA systems have included such a resynthesis pathway by inverting a time-frequency representation (e.g., see [67], [11], [6]). Other representations can also be inverted. For example, Slaney *et al.* [58] describe an approach for inverting the correlogram, which allows sounds to be segregated according to their F0s and then reconstructed.

An advantage of the resynthesis approach is that it allows the use of unmodified ASR systems; this is preferable, because the front-end processing used by CASA systems does not usually provide acoustic features that are suitable for training a conventional ASR system. However, the approach has met with limited success. For example, the system described by Weintraub paired CASA with a speaker-independent continuous-digit-recognition system, and attempted to recognise utterances simultaneously spoken by a male and female speaker. A modest improvement in recognition accuracy was obtained for the (dominant) male voice, but performance for the female speaker actually fell as a result of CASA processing.

A further criticism of the resynthesis approach is that it embodies a very weak link between the CASA and ASR systems; given the important role of schema-driven grouping, one would expect that a tighter integration of CASA and speech models would be beneficial. Ellis [23] has addressed this issue by integrating a speech recogniser into his prediction-driven CASA architecture. When presented with an acoustic mixture, his system attempts to interpret it as speech. Following decoding by the ASR system, an estimate of the speech component of the mixture is used to determine the characteristics of the remaining (nonspeech) signal features. In turn, the estimate of the nonspeech components can be used to re-estimate the speech. Hence, an iterative cycle of estimation and re-estimation develops, which finally converges on an explanation of the acoustic mixture in terms of the speech and nonspeech components present.

Techniques such as correlogram inversion [58] attempt to reconstruct areas of the speech spectrum that have been obliterated by an interfering noise. An alternative approach is to identify those time-frequency regions that are missing or considered unreliable, and treat them differently during the decoding stage of ASR. Specifically, Cooke *et al.* [13] have proposed a *missing feature* approach to ASR which links closely with CASA. In their approach, the observed acoustic features \mathbf{X} are partitioned into two sets, X_r and X_u , which correspond to reliable and unreliable features respectively. Using a modified continuous-density HMM (CDHMM), the authors show that it is possible to impute the values of the missing features X_r . Alternatively, the maximum a posteriori estimate of the speech class can be found as given in (16.10), by replacing the likelihood $P(\mathbf{X}|W)$ with the marginal distribution $P(X_r|W)$. Furthermore, a ‘bounded marginalisation’ approach may be used in which the values of the missing features are constrained to lie within a certain range. For example, the value of a spectral feature must lie between zero and the observed spectral energy.

In practice, a missing feature recogniser is provided with a set of acoustic features and a time-frequency mask, which is typically obtained from a CASA system. The mask may be binary (in which case each time-frequency region is regarded as either reliable or unreliable) or real-valued. In the latter case, each mask value may be interpreted as the probability that the corresponding time-frequency region is reliable.

A number of workers have described ASR systems that combine a missing feature speech recogniser with CASA-based mask estimation. For example, Roman *et al.* [52] employ binaural cues to estimate the ideal binary mask for a target speaker which is spatially separated from an interfering sound source. A large improvement in recognition accuracy was obtained compared to a conventional ASR system, for a connected digit recognition task. Similar results were obtained by Palomäki *et al.* [47], also using a binaural model and missing feature ASR system. Their system computes a binary mask by examining the cross-correlation functions in each time-frequency region. Their

system has been evaluated in the presence of moderate reverberation and obtains substantial ASR improvements.

Neural oscillator frameworks for CASA represent an ideal front-end for missing feature ASR systems, because the activity of oscillators arranged in a time-frequency grid can be directly interpreted as a mask. Brown *et al.* [8] describe an oscillator-based CASA system which segregates speech from interfering noise using F0 information (derived from a correlogram). Additionally, unpitched interference is removed by noise estimation and cancellation; oscillators are deactivated (and hence give rise to a mask value of zero) if they correspond to acoustic components that lie below the noise floor. The authors report good performance on a digit recognition task at low SNRs, but note that unvoiced regions of speech are not represented in the oscillator array; as a result, the performance of their system falls below that of a conventional ASR system at high SNRs.

We also note that mask estimation can be achieved in a purely top-down manner. Roweis [54] describes a technique for estimating binary masks using an unsupervised learning method. Specifically, speaker-dependent HMMs are trained on the speech of isolated talkers, and then combined into a factorial HMM (FHMM). The latter consists of two Markov chains which evolve independently. Given a mixture of two utterances, the underlying state sequence in the FHMM is inferred and the output predictions for each Markov chain are computed. A binary mask is then determined by comparing the relative values of these output predictions.

The missing feature approach achieves a tighter integration between CASA and ASR, but still embodies a unidirectional flow of information from the front-end to the recogniser. However, Barker *et al.* [1] report a further development of the missing feature technique which accommodates data-driven and schema-driven processing within a common framework; the so-called *multisource decoder*. In this approach, it is assumed that the observed features \mathbf{Y} represent a mixture of speech and interfering sound sources. The goal is therefore to find the word sequence W and segregation mask S which jointly maximise the posterior probability,

$$\hat{W}, \hat{S} = \underset{W, S}{\operatorname{argmax}} P(W, S | \mathbf{Y}) \quad (16.11)$$

Barker *et al.* show that $P(W, S | \mathbf{Y})$ can be written in terms of the speech features \mathbf{X} (which are now considered to be unobserved) by integrating over their possible values, giving

$$P(W, S | \mathbf{Y}) = P(W) \left(\int P(\mathbf{X} | W) \frac{P(\mathbf{X} | S, \mathbf{Y})}{P(\mathbf{X})} d\mathbf{X} \right) P(S | \mathbf{Y}) \quad (16.12)$$

As before, $P(W)$ and $P(\mathbf{X} | W)$ in (16.12) represent the language model and acoustic model respectively. However, two new terms are introduced. $P(S | \mathbf{Y})$ is a segregation model, which describes the probability of a particular mask

S given the observed features \mathbf{Y} , but independent of the word hypothesis W . Such information can be obtained from a data-driven CASA system. The remaining term $P(\mathbf{X}|S, \mathbf{Y})/P(\mathbf{X})$ is a likelihood weighting factor. Most importantly, the maximisation in (16.11) occurs over both W and S so that both schema-driven and data-driven information are incorporated in the search.

Barker *et al.* derive an efficient search technique for evaluating (16.12) within a CDHMM system, and test the decoder on a noise-corrupted connected digit task. Segregation masks were obtained using a simple spectral subtraction approach. A reduction in word error rate of about 25% was obtained, relative to a conventional ASR system. The authors predict that further performance gains can be achieved by using CASA processing to estimate the masks.

Yet another way of integrating CASA and speech recognition is to use speech schemas triggered by recognition to restore speech which has been masked by noise. Specifically, Srinivasan and Wang propose a schema-based model for phonemic restoration [60]. Their model estimates reliable time-frequency regions and feeds them to a missing feature recogniser. Successful recognition activates a word template, which is then dynamically time warped to the noisy word so as to restore the speech frames corresponding to the noisy portion of the word. Unlike earlier data-driven efforts, their model can restore both voiced and unvoiced phonemes with a high degree of naturalness.

16.5 CASA Compared to ICA

CASA and ICA differ somewhat in their approaches to speech separation; here, we consider some of the differences and also comment on the possibility of harmonising the two approaches.

Broadly, CASA and ICA differ in a number of respects. For example, CASA emphasises the role of intermediate signal representations such as the correlogram, whereas ICA usually operates directly on the sampled acoustic signal. Likewise, CASA algorithms exploit continuity in time and frequency, whereas ICA does not. The performance profile of ICA also differs substantially from that of human listeners. For instance, ICA typically aims to segregate every source signal from a mixture, whereas human listeners perform figure/ground segregation. Similarly, CASA systems – which are motivated by human performance – often aim to separate a target speaker from the acoustic background rather than completely demix the input (e.g., [66], [40], [8]).

A direct comparison of CASA and ICA was reported by van der Kouwe *et al.* [63]. They compared the performance of Wang and Brown's CASA system [66] with two schemes for ICA, one of which was the fourth-order JADE method [9]. The algorithms were evaluated on Cooke's [11] corpus of speech and noise mixtures, and performance was expressed in terms of the gain in SNR obtained. It was found that the CASA and ICA algorithms

performed well under very different conditions. In general, CASA techniques require that the acoustic mixture exhibits well-defined regions in the time-frequency plane which correspond to one or more sound sources. Hence, the performance of the CASA system was best in conditions in which the interferer was tonal or locally narrowband. The JADE algorithm did not perform as well in these conditions, presumably because the narrowband interferers yielded poor higher-order statistics. On the other hand, the CASA system performed poorly in conditions where there was substantial spectral overlap between the speech and interferer. Again the situation for JADE was the opposite; it performed particularly well with broadband interferers (such as speech and random noise), which contain rich higher order joint statistics.

It should be noted that comparison of CASA and ICA is frustrated by the lack of a suitable corpus. The speech and noise mixtures employed by van der Kouwe *et al.* were not ideal, because the mixing process was constant and linear, the mixing matrix was far from singular, there were two mixtures and two sources, and source signals were perfectly temporally aligned in both mixtures. Such conditions meet all of the requirements for ICA (except for statistical independence of the sources), but are not representative of mixtures recorded in real acoustic environments. On the other hand, the corpus was designed to present a challenging test for CASA systems [11], which do not have such requirements. Clearly, further comparison of CASA and ICA techniques would be facilitated by the availability of a corpus that was designed for evaluating both approaches.

Although CASA and ICA differ in their approaches, there are some similarities between them. For example, de Cheveigné [14] notes the similarity between frequency-domain ICA and equalisation-cancellation models of binaural signal detection. Also, there are possibilities for combining the two approaches [59]. Yilmaz and Rickard [71] describe an approach for separating speech mixtures via the blind estimation of time-frequency masks, which is closely related to the system of Roman *et al.* [52]. Such an approach could be integrated with CASA systems that use a similar time-frequency representation (e.g., [7], [66], [29]). Another example of the combination of ICA and CASA technique is provided by Rutkowski *et al.* [55], who describe a system in which ICA is applied to each frequency channel of a correlogram. The extracted signals in each channel that have a periodic structure are used to reconstruct a time-domain waveform using correlogram inversion [58], whereas the remaining noisy signals are discarded. The authors report good performance for the separation of two sources recorded in a reverberant room, which exceeds the performance expected using CASA or ICA alone.

16.6 Challenges for CASA

In this penultimate section, we briefly review some of the challenges that remain for CASA, and make suggestions for further work.

Evaluation is an important issue for CASA that requires further thought. Research in ASR has undoubtedly benefitted from the adoption of standard metrics and evaluation tasks for comparing performance, such as those introduced by the US National Institute of Standards and Technology (NIST). The situation in CASA is very different; workers rarely compare their work on the same corpus and use a variety of performance metrics. The latter include comparisons of intermediate auditory representations [11], various metrics related to SNR [7], [66] and ASR performance using conventional or ‘missing feature’ speech recognisers [67,52,47]. Ellis [22] argues that the CASA research community should standardise on an evaluation domain that is relevant to a real-world problem (such as acoustic analysis of multi-party meetings), and that the performance of CASA systems should be judged against human performance on the same task.

On a related point, CASA is informed and motivated by the psychophysical literature on ASA (and to a lesser extent, the physiological literature). However, if CASA systems are ‘models’ of human function in a true sense, then they should be able to generate hypotheses that can be tested by further psychophysical experimentation. In fact, there is currently little evidence of such synergy occurring. A notable exception is the work of Cooke [10], who has proposed a ‘glimpsing’ model of human speech perception based on insights gained from his missing feature approach to ASR.

Most work in CASA assumes that sound sources remain in fixed positions for the duration of an input signal. This is not representative of real-world environments, and dealing with moving sound sources remains a challenging research issue. Early work on this problem is reported by Roman and Wang [51], who describe a binaural model based on the same principles as the multi-pitch tracking algorithm of Wu *et al.* [69]. Following auditory filtering and cross-correlation, an HMM is used to form continuous location tracks and estimate the number of sound sources present. Their approach performs well; for example, it is able to simultaneously track two sound sources whose trajectories cross in space.

Another interesting area for further study is the combination of CASA algorithms with established signal processing techniques. For example, Drake *et al.* [19] describes a two-stage algorithm for CASA-enhanced beamforming (CASA-EB). In the first stage of her system, the output from each channel of an auditory model is processed by a beamformer and mapped to a three dimensional space with dimensions of frequency, time and arrival angle. In the second stage, acoustic components are grouped according to F0 and estimated location. Drake *et al.* demonstrate that in most conditions, the performance of CASA-EB is superior to that of monaural CASA or beamforming alone.

The role of attention in CASA has largely been overlooked in computational studies, and merits further work. A model of auditory attention might allow a single source to be tracked in a changing acoustic environment, or allow the most salient source to be extracted from a mixture. Preliminary

work in this regard has been reported by Wrigley and Brown [70]. In their model, a network of neural oscillators performs stream segregation using a principle of oscillatory correlation. A weighting is given to specific frequency regions using a Gaussian-shaped function, which determines the connection weights between oscillators and an attentional unit. When the activity of the attentional unit is synchronised with a group of oscillators, the corresponding acoustic features are held to be in the attentional foreground. The authors have demonstrated the ability of their model to explain psychophysical findings related to the perception of tonal stimuli, but the model remains to be tested with complex stimuli such as speech.

Another challenge for CASA is the monaural separation of unvoiced speech; this issue has received much less attention than the problem of segregating voiced speech using pitch cues. Recently, Hu and Wang [28] have described a system for separating stop consonants from background interference. Stops generally consist of a weak closure and a subsequent burst, which is usually unvoiced and cannot therefore be separated from interference on the basis of pitch. Instead, Hu and Wang's system identifies stops by detecting onsets in the average rate response at the output of each channel of an auditory filterbank. If a significant onset has occurred, it is classified by a Bayesian decision rule on the basis of its spectral shape, intensity and decay time in order to determine whether it corresponds to a stop consonant (as opposed to another impulsive non-speech sound). The authors evaluate their algorithm on a task involving the detection of stop consonants mixed with several types of environmental noise. The system performs respectably when the SNR is high (20 dB or above) but detection performance falls rapidly at lower SNRs. However, the number of confusions (interfering signals which are erroneously identified as stops) remains relatively low, even at 0 dB SNR.

Finally, relatively few workers have evaluated their CASA systems in reverberant conditions, and still fewer have included mechanisms that deal specifically with reverberated input. Exceptions include the binaural model of Palomäki *et al.* [47], which employs a simple model of the 'precedence effect' to remove echoes due to room reverberation. Evaluation of sound separation systems in reverberant conditions has also been undertaken by Kollmeier and Koch [34], Shamsoddini and Denbigh [56] and Roman and Wang [53]. Dealing with reverberation in single-microphone recordings is a particularly challenging issue. Recently, Wu and Wang [68] have proposed an approach to this problem which is based on a two-stage algorithm. In the first stage, a pitch-based metric is used to estimate the reverberation time; this is based on the author's previous multipitch tracking algorithm [69]. In the second stage, an enhancement method estimates and subtracts the acoustic energy due to echoes. An novel approach to the problem is also proposed by Palomäki *et al.* [48]. They describe a system for recognition of reverberated speech in which a time-frequency 'reverberation mask' is estimated for use with a missing feature ASR system. Elements that are selected in the mask correspond

to time-frequency regions that are relatively uncontaminated by reverberation. Because their system is based on a mask representation, it would be relatively straightforward to combine it with other CASA algorithms.

16.7 Conclusions

In summary, CASA aims to replicate the perceptual processes by which human listeners segregate simultaneous sounds. It is a growing research area, which is attracting the interest of workers in the machine learning community as well as those in the signal processing and computational modelling communities. Much progress has been made in CASA-based speech separation in the last few years.

Given the topic of this volume, we conclude with some further comments on the differences between CASA and ICA. Firstly, we note that speech segregation need not require the resynthesis of a high-quality speech signal, as is assumed in most ICA studies. If the goal is robust ASR, then only a time-frequency mask and its corresponding acoustic features are needed. Secondly, we have indicated that there are some prospects for marrying the CASA and ICA approaches. For example, ICA can be used to find the independent components in each channel of an auditory filterbank, or in mid-level auditory representations such as the correlogram.

CASA is motivated by an account of auditory perception; indeed, the term ‘model’ is frequently used to describe CASA systems. We believe that adherence to the general principles of auditory processing is likely to give rise to CASA systems that make fewer assumptions than those based on ICA, and we are hopeful that this will translate into superior performance in real-world acoustic environments.

Acknowledgments

GJB was supported by EPSRC grant GR/R47400/01. DLW was supported in part by grants from the AFOSR, AFRL/IF and NSF.

References

1. J. P. Barker, M. P. Cooke, and D. P. W. Ellis, “Decoding speech in the presence of other sources,” in *Speech Communication*, 2004, in press.
2. F. Berthommier and G. F. Meyer, “Source separation by a functional model of amplitude demodulation,” in *Proc. EUROSPEECH*, 1995, vol. 4, pp. 135–138.
3. F. Berthommier and G. F. Meyer, “Improving amplitude modulation maps for F0-dependent segregation of harmonic sounds,” in *Proc. EUROSPEECH*, 1997, vol. 5, pp. 2483–2486.
4. M. Bodden, “Modelling human sound-source localization and the cocktail party effect,” in *Acta Acustica*, 1993, vol. 1, pp. 43–55.

5. A. S. Bregman, *Auditory scene analysis*. MIT Press, Cambridge MA, 1990.
6. G. J. Brown, *Computational auditory scene analysis: A representational approach*. Ph. D. Thesis, University of Sheffield, 1992.
7. G. J. Brown and M. P. Cooke, "Computational auditory scene analysis," in *Computer Speech and Language*, 1994, vol. 8, pp. 297–336.
8. G. J. Brown, D. L. Wang and J. Barker, "A neural oscillator sound separator for missing data speech recognition," in *Proc. IJCNN*, 2001, vol. 4, pp. 2907–2912.
9. J. F. Cardoso, "High-order contrasts for independent component analysis," in *Neural Computation*, 1999, vol. 11, pp. 157–192.
10. M. P. Cooke, "Making sense of everyday speech: A glimpsing account," in *Speech Separation by Humans and Machines*, edited by P. Divenyi, Springer, New York, 2004.
11. M. P. Cooke, *Modelling auditory processing and organisation*. Cambridge University Press, Cambridge, UK, 1993.
12. M. P. Cooke, G. J. Brown, M. D. Crawford and P. Green, "Computational auditory scene analysis: Listening to several things at once," in *Endeavour*, 1993, vol. 17 (4), pp. 186–190.
13. M. P. Cooke, P. Green, L. Josifovski and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," in *Speech Communication*, 2001, vol. 34, pp. 267–285.
14. A. de Cheveigné, "The cancellation principle in acoustic scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi, Springer, New York, 2004.
15. A. de Cheveigné, "Cancellation model of pitch perception," in *J. Acoust. Soc. Am.*, 1998, vol. 103, pp. 1261–1271.
16. A. de Cheveigné and H. Kawahara, "Multiple period estimation and pitch perception model," in *Speech Communication*, 1999, vol. 27, pp. 175–185.
17. C. J. Darwin and R. P. Carlyon, "Auditory grouping," in *Hearing*, edited by B. C. J. Moore, Academic Press, pp. 387–424, 1995.
18. P. N. Denbigh and J. Zhao, "Pitch extraction and separation of overlapping speech," in *Speech Communication*, 1992, vol. 11, pp. 119–125.
19. L. A. Drake, A. Katsaggelos, J. C. Rutledge, and J. Zhang, "Sound source separation via computational auditory scene analysis-enhanced beamforming," in *Proceedings of the IEEE Sensor Array and Multichannel Signal Processing Workshop*, Rosslyn, VA, 2002.
20. N. I. Durlach, "Note on the equalization and cancellation theory of binaural masking level differences," in *J. Acoust. Soc. Am.*, 1960, vol. 32 (8), pp. 1075–1076.
21. D. P. W. Ellis, *Prediction-driven computational auditory scene analysis*. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, M.I.T., 1996.
22. D. P. W. Ellis, "Evaluating speech separation systems", in *Speech Separation by Humans and Machines*, edited by P. Divenyi, Springer, New York, 2004.
23. D. P. W. Ellis, "Using knowledge to organize sound: The prediction-driven approach to computational auditory scene analysis, and its application to speech/nonspeech mixtures," in *Speech Communication*, 1998, vol. 27, pp. 281–298.
24. D. J. Godsmark and G. J. Brown, "A blackboard architecture for computational auditory scene analysis," in *Speech Communication*, 1999, vol. 27, pp. 351–366.

25. J. G. Harris, C. J. Pu, and J. C. Principe, "A monaural cue sound localizer," in *Analog Integrated Circuits and Signal Processing*, 2000, vol. 23, pp. 163–172.
26. M. J. Hewitt and R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function," in *J. Acoust. Soc. Am.*, 1991, vol. 90 (2), pp. 904–917.
27. G. Hu and D. L. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 79–82.
28. G. Hu and D. L. Wang, "Separation of stop consonants," in *Proc. IEEE ICASSP*, 2003, vol. 2, pp. 749–752.
29. G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," in *IEEE Trans. Neural Networks*, 2004, vol. 15 (5), pp. 1135–1150.
30. L. A. Jeffress, "A place theory of sound localization," in *Journal of Comparative and Physiological Psychology*, 1948, vol. 41, pp. 35–39.
31. A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in *Proc. IEEE ICASSP*, 2000, pp. 2985–2988.
32. M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *Proc. IEEE ICASSP*, 1999, vol. 2, pp. 929–932.
33. A. P. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," in *IEEE Trans. Speech Audio Proc.*, 2003, vol. 11 (6), pp. 804–816.
34. B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," in *J. Acoust. Soc. Am.*, 1994, vol. 95, pp. 1593–1602.
35. A. Khurshid and S. L. Denham, "A temporal analysis based pitch estimation system for noisy speech with a comparative study of performance of recent systems," in *IEEE Trans. Neural Networks*, 2004, vol. 15 (5), pp. 1112–1124.
36. J. C. R. Licklider, "A duplex theory of pitch perception," in *Experimentia*, 1951, vol. 7, pp. 128–133.
37. C. Liu, B. C. Wheeler, W. D. O'Brien, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," in *J. Acoust. Soc. Am.*, 2000, vol. 108 (4), pp. 1888–1905.
38. C. Liu, B. C. Wheeler, W. D. O'Brien, C. R. Lansing, R. C. Bilger, D. L. Jones, and A. S. Feng, "A two-microphone dual delay-line approach for extraction of a speech sound in the presence of multiple interferers," in *J. Acoust. Soc. Am.*, 2001, vol. 110 (6), pp. 3218–3231.
39. R. F. Lyon, "A computational model of binaural localization and separation," in *Proc. IEEE ICASSP*, 1983, pp. 1148–1151.
40. D. Mellinger, *Event formation and separation in musical sound*. Ph.D. Thesis, Stanford University, 1991.
41. B. C. J. Moore, *An introduction to the psychology of hearing* (5th edition). Academic Press, 2003.
42. T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localisation and its application to speech stream segregation," in *Speech Communication*, 1999, vol. 27, pp. 209–222.

43. J. Nix, M. Kleinschmidt, and V. Hohmann, "Computational auditory scene analysis by using statistics of high-dimensional speech dynamics and sound source direction," in *Proc. EUROSPEECH*, 2003, 1441–1444.
44. H. G. Okuno, T. Nakatani, and T. Kawabata, "A new speech enhancement: Speech stream segregation," in *International Conference on Spoken Language Processing*, 1996, vol. 4, pp. 2356–2359.
45. H. G. Okuno, T. Nakatani, and T. Kawabata, "Listening to two simultaneous speeches," in *Speech Communication*, 1999, vol. 27, pp. 299–310.
46. L. Ottaviani and D. Rocchesso, "Separation of speech signal from complex auditory scenes," in *Proceedings of the Conference on Digital Audio Effects*, 2001, Limerick, Ireland.
47. K. J. Palomäki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," in *Speech Communication*, 2004, vol. 43 (4), pp. 273–398.
48. K. J. Palomäki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," in *Speech Communication*, 2004, vol. 43, pp. 123–142.
49. T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," in *J. Acoust. Soc. Am.*, 1976, vol. 60 (4), pp. 911–918.
50. R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner, Pergamon, Oxford, pp. 429–446, 1992.
51. N. Roman and D. L. Wang, "Binaural tracking of multiple moving sources," in *Proc. IEEE ICASSP*, 2003, vol. 5, pp. 149–152.
52. N. Roman, D. L. Wang and G. J. Brown, "Speech segregation based on sound localization," in *J. Acoust. Soc. Am.*, 2003, vol. 114 (4), pp. 2236–2252.
53. N. Roman and D. L. Wang, "Binaural sound segregation for multisource reverberant environments," in *Proc. IEEE ICASSP*, 2004, vol. 2, pp. 373–376.
54. S. T. Roweis, "One microphone source separation," in *Neural Information Processing Systems*, 2000, vol. 13, pp. 793–799.
55. T. Rutkowski, A. Cichocki, and A. K. Barros, "Speech enhancement from interfering sounds using CASA techniques and blind source separation", in *3rd International Conference on Independent Component Analysis and Blind Signal Separation*, 2001, San Diego, California, pp. 728–733.
56. A. Shamsoddini and P. N. Denbigh, "A sound segregation algorithm for reverberant conditions," in *Speech Communication*, 2001, vol. 33, pp. 179–196.
57. M. Slaney and R. F. Lyon, "A perceptual pitch detector," in *Proc. IEEE ICASSP*, 1990, vol. 1, pp. 357–360.
58. M. Slaney, D. Naar, and R. F. Lyon, "Auditory model inversion for sound separation," in *Proc. IEEE ICASSP*, 1994, pp. 77–80.
59. P. Smaragdis, *Redundancy reduction for computational audition, a unifying approach*. Ph.D. Thesis, Program in Media Arts and Sciences, M.I.T., 1994.
60. S. Srinivasan and D. L. Wang, "A schema-based model for phonemic restoration," in *Speech Communication*, 2004, in press.
61. H. W. Strube, "Separation of several speakers recorded by two microphones (cocktail-party processing)," in *Signal Processing*, 1981, vol. 3 (4), pp. 355–364.
62. T. Tolonen and M. Karjalainen, "A computationally efficient multi-pitch analysis model," in *IEEE Trans. Speech Audio Proc.*, 2000, vol. 8 (6), pp. 708–716.

63. A. J. W. van der Kouwe, D. L. Wang and G. J. Brown, "A comparison of auditory and blind separation techniques for speech segregation," in *IEEE Trans. Speech Audio Proc.*, 2001, vol. 9 (3), pp. 189–195.
64. D. L. Wang, "Primitive auditory segregation based on oscillatory correlation," in *Cognitive Science*, 1996, vol. 20, pp. 409–456.
65. D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi, Springer, New York, 2004.
66. D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," in *IEEE Trans. Neural Networks*, 1999, vol. 10 (3), pp. 684–697.
67. M. Weintraub, *A theory and computational model of monaural auditory sound separation*. Ph. D. Thesis, Stanford University, 1985.
68. M. Wu and D. L. Wang, "A one-microphone algorithm for reverberant speech enhancement," in *Proc. IEEE ICASSP*, 2003, vol. 1, pp. 844–847.
69. M. Wu, D. L. Wang and G. J. Brown, "A multipitch tracking algorithm for noisy speech," in *IEEE Trans. Speech Audio Proc.*, 2003, vol. 11 (3), pp. 229–241.
70. S. N. Wrigley and G. J. Brown, "A computational model of auditory selective attention," in *IEEE Trans. Neural Networks*, 2004, vol. 15 (5), pp. 1151–1163.
71. O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," in *IEEE Trans. Signal Processing*, 2004, vol. 52 (7), pp. 1830–1847.