Listen. What do you hear? Perhaps a passing car, distant voices, the hum of a computer or music playing quietly in the background. Regardless, it is likely that you can identify more than one source of sound. This illustrates a property of human hearing that is so familiar we usually take it for granted: we are able to distinguish individual sound sources from a complex mixture of sounds reaching our ears. A striking example of this perceptual feat is given by Helmholtz [53] in his 1863 masterpiece on audition:

In the interior of a ball-room . . . we have a number of musical instruments in action, speaking men and women, rustling garments, gliding feet, clinking glasses, and so on . . . a tumbled entanglement [that is] complicated beyond conception. And yet . . . the ear is able to distinguish all the separate constituent parts of this confused whole. ([53], pp. 26–27)

A particularly important aspect of this problem, which is apparent from Helmholtz’s description, is the perception of speech in such complex acoustic environments. Writing in the 1950s, Cherry [24, 25] termed this the “cocktail party problem”:

One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others . . . we may call it “the cocktail party problem.” No machine has yet been constructed to do just that. ([25], p. 280)

Cherry’s assessment of machine performance is as pertinent today as it was half a century ago. How indeed might one construct a device that can segregate the mixture of sounds emanating from a cocktail party, or from any other acoustic environment, with a performance that matches that of human listeners? This challenging problem is the topic of this book.

Insight into the processes underlying the perceptual separation of sound sources
has come from several decades of psychophysical research. In 1990, Bregman [13] published a coherent account of this work. He contends that there are many similarities between audition and vision. When we view a visual scene, the edges, textures and colors are analyzed and interpreted as perceptual wholes (e.g., a face or a chair). Similarly, Bregman argues that the sound reaching our ears is subjected to an auditory scene analysis (ASA). Conceptually, this may be regarded as a two-stage process. In the first stage (which has been called segmentation), the acoustic input is decomposed into a collection of local time-frequency regions (which we term "segments"). The second stage (grouping) combines segments that are likely to have arisen from the same environmental source into a perceptual structure, which is termed a stream. Bregman’s account therefore makes an important distinction between the acoustic source (such as the plucking of a guitar string) and the corresponding perceptual stream (the mental representation of a guitar being played).

By providing a coherent framework for understanding the perceptual organization of sound, Bregman’s book has stimulated much interest in computational studies of ASA. Such studies are motivated in part by the demand for practical sound separation systems, which have many applications including noise-robust automatic speech recognition, hearing prostheses, and automatic music transcription. This emerging field has become known as computational auditory scene analysis (CAS). This book aims to provide a comprehensive account of the state-of-the-art in CASA, in terms of the underlying principles, the algorithms and system architectures that are employed, and the likely applications of this new technology.

In this introductory chapter, we review general principles of auditory scene analysis in listeners and machines. Section 1.1 briefly describes the structure and function of the human auditory system, and then discusses the auditory organization of simple stimuli (such as sequences of tones). We then consider the performance of human listeners in cocktail party scenarios, in which speech is contaminated by noise or other voices, and review the grouping principles that are thought to underlie ASA. Section 1.2 defines CASA and its goal in an attempt to provide a coherent theme for the field. The section also discusses potential applications of CASA technology. Section 1.3 introduces some fundamental concepts of CASA systems, including system architecture and auditory-motivated signal representations. Section 1.4 discusses how to properly evaluate a CASA system. Section 1.5 summarizes other major sound separation approaches, and contrasts them with CASA. Finally, we end the chapter with a historical account of CASA development, and a brief review of the following chapters of this book.

1.1 HUMAN AUDITORY SCENE ANALYSIS

1.1.1 Structure and Function of the Auditory System

Before considering ASA in human listeners, it is useful to have some appreciation of the structure and function of the early stages of auditory processing. The review that follows is necessarily brief; the reader is referred to standard texts such as [108, 68] for further details.

The auditory periphery is a complex transducer that converts sound vibrations into action potentials ("spikes") in the auditory nerve. Broadly, the periphery can be divided into three areas: the outer, middle, and inner ear. The outer ear consists of the external ear (pinna), the ear canal (meatus), and the eardrum (tympanic membrane); the ear canal connects the external ear with the eardrum. The pinnae play a role in spatial hearing, by imposing spectral characteristics on sound that depend on its direction of incidence with respect to the head. Sound traveling down the ear canal causes the eardrum to vibrate. These vibrations are transmitted to the fluid-filled cochlea by the middle ear, in which three tiny bones (ossicles) form a lever, the action of which matches the impedance of air to that of the cochlear fluids.

The cochlea in the inner ear is the organ of hearing. It consists of a coiled, fluid-filled tube that is divided along its length by two membranes, Reissner’s membrane and the basilar membrane. The basilar membrane varies in mass and stiffness along its length, so that different regions of the membrane vibrate at different resonant frequencies. As a result, the basilar membrane exhibits a complex pattern of motion in response to sound-induced movement of the cochlear fluids. More specifically, the response of the basilar membrane to a sinusoidal stimulus is a traveling wave that moves along the length of the cochlea. Importantly, the traveling wave oscillates at the frequency of the stimulus and, therefore, provides a timing code for stimulus frequency. Additionally, the traveling wave reaches its peak amplitude at a location where the stimulus frequency matches the resonant frequency of the basilar membrane. Hence, stimulus frequency is also represented by a place code.

The movements of the basilar membrane are transduced into neural activity by inner hair cells. The hairs at the tops of these cells are displaced when the basilar membrane moves up and down. As a result, the inner hair cells are activated, and this leads to the initiation of action potentials in the spiral ganglion cells whose axons make up the auditory nerve. It should be noted that the auditory nerve encodes a half-wave rectified version of the stimulus, because action potentials are only initiated by the movement of hairs in one direction.

Auditory nerve responses exhibit a number of important properties. They respond preferentially to certain frequencies, so that the frequency selectivity originating at the basilar membrane is maintained. If the stimulus frequency is low, the fibers respond to individual vibrations of the stimulating waveform, exhibiting the so-called "phase locking" phenomenon. Auditory nerve fibers fire spontaneously in the absence of a stimulus. The firing rate of a fiber is related to stimulus level by a sigmoidal function, so that the neural response exhibits level compression and saturation effects. Also, auditory nerve fibers adapt to a steady stimulus; they respond vigorously at stimulus onset, and then the firing rate declines to a steady state. Finally, there is a refractory period for a brief time after the offset of the stimulus, during which the firing rate falls below the spontaneous level.

Beyond the auditory periphery, the pathway leading to the auditory cortex passes
through four neural structures: cochlear nucleus, superior olive, inferior colliculus, and medial geniculate nucleus. Neurons throughout these higher centers of the auditory pathway and the auditory cortex appear to be tuned to particular stimulus features, such as periodicity, sound intensity, amplitude and frequency modulation, and interaural time and intensity differences.

Although it is clear that human listeners are able to solve the ASA problem from eardrum vibrations in response to sound waves, relatively little is known about the physiological level about the way in which the higher auditory centers perform ASA. However, insights have also come from perceptual experiments; it is these that we consider next.

### 1.1.2 Perceptual Organization of Simple Stimuli

Much of the evidence for ASA has come from perceptual experiments with simple stimuli. A typical example is the sequence of alternating high- and low-frequency tones shown in Fig. 1.1, due to Noorden [126]. When this sequence is presented to listeners at a slow rate (e.g., when the time between tone onsets is 150 ms or so), the perceptual organization of the tones depends upon the frequency difference between them. If the frequency difference is small (less than 4 semitones or so, see Fig. 1.1, panel A) then listeners perceive a single stream consisting of a sequence of high- and low-pitched tones with a galloping rhythm. However, when the high and low tones are separated in frequency by more than 12 semitones or so (panel B), the galloping rhythm is no longer heard. Instead, the sequence segregates into two streams, each of which consists of tones with the same pitch that are equally spaced in time. This phenomenon is known as streaming. At intermediate frequency separations, listeners may hear either organization, and can switch between one or two streams with conscious effort.

![Figure 1.1](image1.png)

**Figure 1.1** Perceptual organization of alternating tone sequences. (A) A tone sequence in which there is a small difference in frequency between the high and low tones is perceived as a perceptual whole (i.e., a single stream). (B) A larger difference in frequency causes the high and low tones to segregate into different streams.

The streaming phenomenon is also dependent on the rate of presentation; at faster rates, the range of ambiguity is reduced, and streaming can be induced by relatively small differences in frequency. Indeed, when the rate of presentation is fast and there is a large frequency separation between the high and low tones, listeners cannot hear a single stream, even with conscious effort. Bregman [13] argues that this effect is due to the existence of automatic, primitive processes of auditory organization. We will expand on this important point in Section 1.1.4. However, before doing so we return to the issue of “cocktail party” listening, in which the stimulus reaching the ears is considerably more complex than the tone sequences discussed above.

### 1.1.3 Perceptual Segregation of Speech from Other Sounds

Although Helmholtz marveled at the ability of human hearing in cluttered acoustic environments, he was also very much aware of its limitations. Specifically, he noted that in order for a specific sound to be heard out from the whole, it must not be “too much overpowered by the mere loudness of the others” ([53], p. 25). In essence, Helmholtz’s comment refers to the phenomenon of masking, in which the threshold of audibility for one sound is raised by the presence of another sound (which is called the masker). So how well can the auditory system segregate, and hence recognize, speech in the presence of masking sounds?

One way to quantify speech intelligibility in noise is the speech reception threshold (SRT), which is the signal-to-noise ratio (SNR) measured in decibels (dB) required for a 50% intelligibility score. Steeneken [122] has investigated the intelligibility of different kinds of speech material when presented in a background of speech-shaped noise (i.e., stationary noise with a spectrum similar to the long-term speech spectrum). His data are shown in Fig. 1.2, and demonstrate that the speech intelligibility score varies with the speech material used. Nonsense consonant-vowel-consonant (CVC) syllables are the most susceptible to noise because they lack contextual cues, having a SRT of about −4 dB. The most intelligible speech materials are spoken digits and letters of the alphabet, which have a SRT of about −10 dB. To put these figures into context, Bronkhorst and Plomp [14] report that for “cocktail party”-like situations in which all voices are equally loud, speech is intelligible for normal-hearing listeners even when as many as six interfering talkers are present. Assuming that the individual utterances are uncorrelated, this situation amounts to a SNR of about −7.8 dB.

The nature of the interference also has a strong influence on the SRT. Miller [93] reviewed the masking of speech by a variety of tones, broadband noises, and other voices. Subjects were tested for their word intelligibility scores, and the results are shown in Fig. 1.3. In general, tones are less effective maskers than broadband noises. For example, speech is intelligible even when corrupted by a complex tone glide that is 20 dB more intense (pure tones are even weaker maskers). Broadband noise is the most effective speech masker, and the corresponding SRT is about 2 dB. When the masker consists of other voices, the SRT varies within a considerable range depending on how many talkers are present. As shown in Fig. 1.3, the SRT is
Speech intelligibility score with respect to SNR for different speech materials (adapted with permission from [122]). The 50% intelligibility score is indicated by the dashed line. A diagnostic rhyme test determines an intelligibility score from a limited number of alternative words that differ on a single articulatory feature. The interference is speech-shaped noise. “PB” stands for “phonetically balanced.”

Figure 1.2 Speech intelligibility score with respect to SNR for different types of interference (redrawn from [93]). The level of the target speech is fixed at 95 dB. The 50% intelligibility is indicated by the dashed line. The pitch frequency of the complex tone rises slowly and drops suddenly in a periodic manner. The broadband noise covers the spectrum from 20 Hz to 4 kHz. For speech interference, data are shown for 1, 2, and 8 competing speakers.

Word intelligibility score (%)

Figure 1.3 Speech intelligibility score with respect to SNR for different speech interference. The SRT stays about the same (around -1 dB) when the masker contains four or more voices.

The relative ease of speech segregation with a single competing talker is usually attributed to the ability of listeners to “glimpse” the target voice during gaps in the masker [94, 27, 2]. More specifically, speech energy is sparsely distributed in the time-frequency plane, and, hence, gaps occur in the spectrum of the masker during which listeners can obtain an uncorrupted estimate of the spectrum of the target voice. The opportunity for glimpsing is much reduced. However, even with a single interfering voice, speech intelligibility is worse when the competing talker is of the same sex than when it is of a different sex, and it is worst when the same talker is used both as the target and interference [22]. Although glimpsing is likely to play a role, it may not adequately explain the large intelligibility differences under such conditions [121]. Additionally, auditory scene analysis may contribute to a glimpsing process by identifying time-frequency regions that are dominated by a single voice.

It is worth noting that the curves shown in Figs. 1.2 and 1.3 are steep in the region of the SRT. To that even a small gain in SNR leads to an appreciable increase in intelligibility (e.g., a SNR gain of 1 dB near the SRT leads to an increase in intelligibility of 5–10%, depending on the interference material). From a different perspective, it has been estimated that hearing-impaired listeners typically need a SNR that is 5–10 dB higher than that of normal listeners in order to achieve the same intelligibility score [110, 111, 1]. This suggests that consistent improvements in SNR of just a few decibels can yield significant improvements in speech intelligibility.

All of the data discussed above concern mixtures of speech and noise that are created in the laboratory and presented to listeners over a single audio channel. However, in natural “cocktail party” environments, listeners receive an input from both ears and must also contend with room reverberation. What is the effect of these factors?

When target speech and noise intrusions are presented from different spatial locations, binaural hearing can lead to a significant reduction in the SRT [14, 38]. At least two mechanisms appear to contribute to this “binaural advantage.” First, when sounds are presented at different locations, the SNR at one ear will be higher than the SNR at the other. Hence, listeners may gain some benefit by selectively attending to the ear in which the SNR is greatest. Second, interaural interaction enables
location-based organization of the auditory scene, a topic discussed in Chapters 5 and 6.

Although speech perception in quiet is robust to room reverberation [97, 7], speech perception in noise is sensitive to reverberation effects. The studies by Plomp [109] and Culling et al. [30] demonstrate a significant increase in the SRT when speech is presented together with an interfering sound in reverberant conditions. For example, Culling et al. report a 5 dB increase in the SRT for naturally intonated speech presented with a reverberation time of 400 ms, compared to anechoic conditions (monotone speech fares a little better). The same study also finds that the benefit of spatial separation disappears in reverberant conditions. The effects of reverberation on speech perception and auditory grouping cues are addressed in Chapter 7.

1.1.4 Perceptual Mechanisms

What are the perceptual mechanisms that underlie auditory organization in human listeners? Bregman’s book, Auditory Scene Analysis, gives the most systematic and comprehensive treatise on this question to date [13]. As explained in the introduction to the book, Bregman’s conceptual framework for ASA amounts to an analysis–synthesis process; the acoustic scene is decomposed into a collection of segments, which are subsequently grouped to form coherent streams. Such grouping processes may operate both simultaneously (in which concurrent segments are integrated) or sequentially (in which segments are grouped across time). Bregman also makes a further distinction between primitive grouping and schema-based grouping. The former is regarded as an innate and bottom-up process, which relies on the intrinsic structure of environmental sounds. Schema-based grouping refers to the fact that auditory features belonging to the same learned pattern, such as a syllable, tend to be bound together. It is therefore a top-down process, and based on prior knowledge.

According to Bregman, primitive grouping is governed by mechanisms that are analogous to those proposed by the Gestalt psychologists in relation to visual perception (for a review, see [102]). If we view the acoustic signal in the form of a time-frequency scene (i.e., like a spectrogram—see Figs. 1.1 and 1.4), then the major primitive grouping principles can be summarized as follows (see among others [126, 51, 13, 31, 64, 141, 8, 96]).

- **Proximity in frequency and time.** The closer acoustic components are in frequency, the greater is the tendency to group them into the same stream. This is apparent from the alternating tone sequences shown in Fig. 1.1. Similarly, acoustic components tend to be perceptually grouped if they are close in time; this explains the effect of presentation rate on the streaming phenomenon, as discussed in Section 1.1.2.

- **Periodicity.** A set of acoustic components that are harmonically related (i.e., have frequencies that are integer multiples of the same fundamental frequency) tend to be grouped into the same stream. This is a simultaneous grouping principle.

- **Continuous or smooth transition.** Frequency components tend to be grouped into the same stream if they form a continuous trajectory (the temporal continuity principle) or a discontinuous but smooth trajectory. Similarly, smooth changes in pitch contour, intensity, spatial location, and spectrum tend to be interpreted as the continuation of an existing sound, whereas abrupt changes signify the onset of a new sound source.

- **Onset and offset.** Listeners tend to group frequency components into the same stream if they have the same onset time or, to a lesser extent, the same offset time. This is a kind of simultaneous organization.

- **Amplitude and frequency modulation.** Frequency components that exhibit the same temporal modulation tend to be grouped together into the same stream. This simultaneous grouping principle applies to both amplitude modulation (AM) and frequency modulation (FM).

- **Rhythm.** A sequence of rhythmically related tones tends to be integrated into the same stream. This is a sequential organization cue that applies to events that are separated in time.

- **Common spatial location.** Concurrent sounds that originate from the same location in space tend to be grouped. Also, as noted above, sequential integration is promoted by smoothly varying location for the case of moving sound sources. However, the fact that listeners can separate sounds that originate from the same location, and can segregate acoustic mixtures when listening monaurally, suggests that spatial location may play a secondary role to spectral cues.

Since these organizational principles have largely been derived from the study of simple laboratory stimuli, we should ask whether they are also applicable to complex, real-world sounds such as speech and music. Figure 1.4 illustrates some of the grouping cues that are present in a speech utterance. Both broadband and narrowband spectrograms are displayed in the figure. Prominent features in this example include the temporal continuity of harmonics and formants, harmonicity of voiced speech, onset and offset synchrony, and coherent amplitude modulation. Despite the complexity of the speech signal, the figure suggests that many primitive grouping cues are applicable to the perceptual organization of natural speech. Further discussion of the role of ASA in speech perception can be found in Chapter 9.

Similarly, it is well known that music is subject to auditory organization. Indeed, this fact was exploited by some Baroque composers, who used sequences of rapidly alternating high and low notes, like those shown in panel B of Fig. 1.1, to give the impression that a single instrument was playing two melodic lines (so-called “virtual polyphony”; see [13], p. 464). A further example is the perception of “interleaved melodies” [36, 37]. In this paradigm, the notes of two melodies are interleaved in time, so that adjacent tones are from different melodies. If the pitch ranges of the two melodies do not overlap, the interleaved sequence is perceptually segregated.
into two streams, which correspond to the two melodies. However, if the two melodies have overlapping pitch ranges, listeners are unable to segregate them. Similar effects are found if the notes of the two melodies are distinguished by timbre, rate of attack and decay, or duration [52].

1.2 COMPUTATIONAL AUDITORY SCENE ANALYSIS (CASA)

1.2.1 What Is CASA?

Broadly speaking, CASA may be defined as the study of auditory scene analysis by computational means. To follow the earlier quote from Cherry, one may define the CASA problem as the challenge of constructing a machine system that achieves human performance in ASA. This definition is entirely functional; it makes no reference to underlying mechanisms. In other words, it makes no commitment as to whether CASA should employ the perceptual and neural mechanisms known to be used by the human auditory system.

One way to make CASA more biologically relevant is to limit the scope of investigation to monaural (one-microphone) or binaural (two-microphone) input, because the information available to the auditory system for solving the ASA problem is the sound received at the two ears. Without some constraint on the number and placement of microphones, the CASA problem could be totally circumvented; imagine, say, a solution that deploys a dedicated microphone for each sound source. As discussed in Section 1.5, it is well understood that a microphone array can be constructed to perform beamforming and spatial filtering. If the target source originates from a distinct direction, beamforming is able to segregate the target sound. However, such a solution is based on a principle that cannot be applied in monaural or binaural conditions, and it only works in constrained environments. On the basis of these considerations, we propose a more specific definition of CASA: It is the field of computational study that aims to achieve human performance in ASA by using one or two microphone recordings of the acoustic scene.

As noted above, the notion of CASA does not imply that the human auditory system should be slavishly modeled. However, many workers adopt an approach that is to some extent based on the principles of processing in the human auditory system. As a result, the term CASA has come to be associated with a perceptually motivated approach that is distinct from other approaches to sound separation (see Section 1.5). In practice, the influence of perceptual studies on CASA research varies considerably. Indeed, the following chapters of this book illustrate that CASA systems range from models that explicitly simulate perceptual and physiological data, to sound separation systems that are related to perception only at a very abstract level.

It is sometimes said that the CASA field lacks a common theory. Comparison is often made with automatic speech recognition (ASR), in which the goal is clear and the evaluation metric is common. Although it is true that there is no consensus on the criteria for evaluating CASA systems (see Section 1.4.1), we should bear in
mind that the problem of CASA is considerably broader than the ASR problem. Given the interdisciplinary nature of the field and the diverse range of motivations and applications, it should not be surprising that there are different opinions on the essence of CASA and it will likely remain so in the foreseeable future. One could say the same of computational vision, which has been developed by many more workers for a much longer time. In our opinion, it is healthy to have diversity in the early stages of development of this new field.

1.2.2 What Is the Goal of CASA?

In his influential treatise on computational vision, Marr [85] makes a compelling argument for separating three levels of analysis in order to understand complex information processing systems. The first level—computational theory—is concerned with analyzing the goal of computation and the general processing strategy. The second level—representation and algorithm—is concerned with the representation of the input and output, and the algorithm that transforms from the input representation to the output representation. The third level—implementation—is concerned with how to physically realize the representation and algorithm. Marr states that, “Each of the three levels of description will have its place in the eventual understanding of perceptual information processing” (p. 25).

Marr’s framework suggests a key question—What is the goal of CASA? In order to answer this, we should, as suggested above, put the broader context of auditory perception and ask what purpose perception serves. From an information processing perspective, which is shared by human and machine perception, Gibson [45] considers perceptual systems to be ways of seeking and extracting information about the environment from the sensory input. More specifically, the purpose of vision, according to Marr [85], is to produce a visual description of the environment for the viewer. By analogy with this, we may state that the purpose of audition is to produce an auditory description of the environment for the listener. Two related points are worth noting here. First, perception is a process private to the perceiver, despite the fact that the physical environment may be shared by different perceivers. Second, what perception gives us at a given time is a description of the environment, not the description. That the perceiver constructs only a partial picture of the environment is well illustrated by the perceptual phenomenon of change blindness: a change to an image during a flicker or other interruption often goes undetected, despite the fact that the change is easily seen once the attention of the viewer is directed to it (for example, see [112]). Change blindness is attributed to the limited capacity of attention.

Considering auditory scene analysis more specifically, Bregman [13] states that the goal of ASA is to produce separate streams from the auditory input, such that each stream represents a single sound source in the acoustic environment. To extrapolate from this, we may state that the goal of CASA is to computationally extract individual streams from one or two recordings of an acoustic scene. Note that here and elsewhere in this volume we use the term “stream” to refer both to the perceptual representation of a sound source, and to the representation of a sound source in computer memory.

An analysis of the CASA problem at the computational theory level has led Wang [132] to propose that the goal of CASA should be to estimate an ideal time-frequency (T-F) mask. Consider a time-frequency representation such as the spectrogram shown in Fig. 1.4, in which the frequency axis and time axis are divided into discrete units. An ideal T-F mask is then a binary matrix, whose value is one for a T-F unit in which the target energy is stronger than the total interference energy, and is zero otherwise. Further discussion of this approach is given in Section 1.3.5.

1.2.3 Why CASA?

In addition to the scientific challenge of constructing a “cocktail party processor,” research in CASA has a number of important applications, several of which are listed below.

- **Robust automatic speech and speaker recognition.** Much progress has been made in automatic speech and speaker recognition in recent years. However, the performance of these systems degrades rapidly in the presence of acoustic interference, and is much poorer than human performance [81, 59]. Arguably, the current predicament of recognition systems in real environments is largely due to a preoccupation with the recognition of clean speech. This stands in contrast to work in computer vision, in which the central role of scene (image) analysis was recognized from the beginning [85, 43]. CASA offers to redress this imbalance by providing a perspective in which speech is regarded as just one of many sound sources in a complex acoustic environment.

- **Hearing prostheses.** About 10% of the population suffers from hearing loss, and for most of them hearing aids are the primary means of alleviating the associated deficits. Listeners with hearing loss often have difficulty in understanding speech in noisy environments (i.e., their ability to perform “cocktail party” processing is reduced). In such cases, hearing aids that provide amplification are of little help, since they amplify both the speech and the noise. Noise robustness is regarded as the greatest obstacle in hearing aid design [35, 76], and it is a problem for which CASA could provide a solution.

- **Automatic music transcription.** Automatic music transcription aims to derive a symbolic score (usually in a note-based form) from a musical audio recording. This is a challenging technical problem in its own right, and a solution to the problem also offers the potential for notating ethnic music that has no written form. The problem is similar in some respects to automatic speech recognition, but with the added complication that music contains concurrent instruments that must be segregated before they can be individually transcribed. CASA may be able to provide such a segregation (e.g., see [17, 48]). More broadly, CASA can also contribute to solving the problem of music scene description, which is addressed in Chapter 8.
• **Audio information retrieval.** A huge volume of audio recordings are available in private archives and via the Internet, and a key research problem is how to search the audio content efficiently. Because recordings generally contain mixtures of sound sources, the separation of such mixtures is often a prerequisite for retrieving audio information from sound files.

• **Auditory scene reconstruction.** Once an acoustic mixture has been segregated, an auditory scene can be reconstructed in which acoustic sources are placed at arbitrary locations. For example, an interface may be designed to selectively modify desired acoustic events, or to present different sounds at perceptually distinct locations for enhanced human listening.

• **Contribution to hearing science.** CASA can make a contribution to hearing science, by suggesting computational mechanisms that help to explain how the auditory system solves the ASA problem.

### 1.3 BASICS OF CASA SYSTEMS

In this section, we introduce some of the fundamental ideas associated with the design and implementation of CASA systems. First, we discuss the issue of system architecture—the sequence of processing blocks that comprise a CASA system, and the manner in which information is passed between them. We then discuss some auditory-motivated signal representations (e.g., the correlogram and cross-correlogram), which have frequently been used in CASA systems. The notion of a time-frequency mask is then introduced, and, finally, we discuss strategies for resynthesizing audio waveforms of segregated sound sources.

#### 1.3.1 System Architecture

Figure 1.5 shows a representative architecture of CASA systems, which broadly follows Bregman’s conceptual framework for ASA. In this architecture, a digitally recorded acoustic mixture first undergoes peripheral analysis, giving a time-frequency representation of auditory activity (e.g., a cochleagram; see Section 1.3.2). Acoustic features are then extracted, such as periodicity, onsets, offsets, amplitude modulation, and frequency modulation. Mid-level representations, such as segments or other intermediate descriptions, are then formed using these features. Scene organization takes place on the basis of primitive grouping cues and trained models of individual sound sources (and the background), producing separate streams. Finally, an audio waveform is resynthesized from a separated stream, so that segregation performance can be assessed by listening tests or by metrics that compare the segregated waveform with a ground truth.

#### 1.3.2 Cochleagram

The first stage of processing in a CASA system usually derives a time-frequency representation of the acoustic input. This is often achieved by using a computer model of the peripheral auditory system, in order to provide a frequency analysis that is consistent with the known properties of human frequency selectivity. Computer models of the auditory periphery broadly follow the structure described in Section 1.1.1, with processing components that simulate the outer/middle ear, cochlear frequency selectivity, and transduction by hair cells. Notable integrated models of this type include those of Lyon [82], Seneff [117], Deng and Geisler [34], Patterson et al. [104], and Zhang et al. [142].

When modeling cochlear frequency selectivity, it is usual to adopt a functional approach rather than use a detailed model of basilar membrane mechanics. Generally, the basilar membrane is modeled either as a transmission line (i.e., a cascade of filter sections) [82] or as a filterbank, in which each filter (or “channel”) models the frequency response associated with a particular point on the basilar membrane. Here, we focus on a widely adopted example of the latter approach, which is based on the gammatone filter.

The gammatone filter was popularized by Johannesma [63] as a model for the impulse response function of auditory nerve fibers, as estimated by the reverse correlation of spike patterns (see also [42, 32]). The gammatone is a bandpass filter, whose impulse response $g_{\phi}(t)$ is the product of a gamma function and a tone (hence “gammatone”):

$$g_{\phi}(t) = \prod_{\nu=1}^{N-1} \exp[-2\pi b(f_\nu)] \cos(2\pi f_\nu t + \phi)u(t) \quad (1.1)$$

Here, $N$ is the filter order, $f_\nu$ is the filter center frequency (in Hz), $\phi$ is the phase, and $u(t)$ is the unit step function (i.e., $u(t) = 1$ for $t \geq 0$, and 0 otherwise). The function $b(f_\nu)$ determines the bandwidth for a given center frequency. For $f_\nu/b(f_\nu)$ sufficiently large, a good approximation to the frequency response of the gammatone is given by [55]

$$G(f) \approx \left[1 + \frac{j(f - f_\nu)}{b(f_\nu)}\right]^{-N} \quad (0 < f < \infty) \quad (1.2)$$

![Figure 1.5 System architecture of a typical CASA system.](image-url)
It can therefore be seen that the filter is symmetric about $f_c$ on a linear frequency scale. Patterson et al. [105] show that for $N = 4$, the gammatone filter gives an excellent fit to experimentally derived estimates of human auditory filter shape.

The bandwidth of the gammatone filter is usually set according to measurements of the equivalent rectangular bandwidth (ERB) of human auditory filters [46]. The ERB of a filter is defined as the bandwidth of an ideal rectangular filter that has the same peak gain, and which passes the same total power for a white-noise input. For auditory filters, the ERB may be regarded as a measure of critical bandwidth [96], and a good match to human data is given by

$$\text{ERB}(f) = 24.7 + 0.108f$$  \hspace{1cm} (1.3)$$

Furthermore, it is usual to assume that the filter center frequencies are distributed over frequency in proportion to their bandwidths, which for fourth-order filters is given by [105]

$$b(f) = 1.019 \text{ERB}(f)$$  \hspace{1cm} (1.4)$$

In this regard, it is often convenient to distribute the filter center frequencies on the so-called “ERB-rate” scale. This is a warped frequency scale, similar to the critical-band scale of the human auditory system, on which filter center frequencies are uniformly spaced according to their ERB bandwidth. The ERB-rate scale is an approximately logarithmic function relating frequency to the number of ERBs, $E(f)$, and is given by

$$E(f) = 21.4 \log_{10}(0.00437f + 1)$$  \hspace{1cm} (1.5)$$

Figure 1.6 shows the impulse responses and frequency responses for eight gammatone filters that are equally spaced on the ERB-rate scale. Note that at low frequencies, the filters have narrow bandwidths, and are more closely spaced in frequency (panel B). In the figure, the filters have been normalized to have the same peak gain at their center frequency. In practice, however, the gains can be set in order to simulate the resonances due to the outer and middle ears, which boost sound energy in the 2–4 kHz range. Alternatively, the outer/middle ear transfer function can be imposed on the input signal, prior to frequency analysis, by a simple linear filter.

Panel A of Fig. 1.6 also illustrates that the impulse responses of the low-frequency filters peak at a later time than those of the high-frequency filters, due to their narrow bandwidths. If across-frequency comparisons are to be made in a CASA system (e.g., in order to detect event onsets in different frequency channels) then it may be convenient to phase-compensate the gammatone filter bank, so that the peaks of the impulse responses are aligned. Holdsworth et al. [55] show that this can be achieved by introducing a time lead $t_c = (N - 1)/2\pi b(f)$ to the output of each filter, in order to align the peaks of their envelopes. A further phase correction $\phi = -2\pi f_c t_c$ is needed to align the peak of the envelope of each impulse response with the peak of its fine structure, giving the phase-compensated filter

$$\tilde{g}_f(t) = (t + t_c)^{N-1} \exp[-2\pi b(f)(t + t_c)] \cos(2\pi f_c t) u_c(t)$$  \hspace{1cm} (1.6)$$

where $u_c(t) = 1$ for $t \geq -t_c$, and 0 otherwise. Note that the resulting filter is non-causal.

Efficient digital implementations of the gammatone filter have been proposed by Holdsworth et al. [55], Cooke [28], and Slaney [118]. The output of the filter may be regarded as a measure of basilar membrane displacement, which is subjected to further processing in order to derive a simulation of auditory nerve activity. For reasons of computational efficiency, most CASA systems use a representation of firing rate in the auditory nerve, rather than a spike-based representation. Most simply, this can be obtained by half-wave rectification of the filterbank outputs, followed by a static nonlinearity (such as a square root). A more sophisticated approach is to model adaptation using an automatic gain control (AGC) [117], or to use a model of hair cell transduction (see [54] for a review).

The model of hair cell transduction proposed by Meddis [87, 88, 91] is often paired with the gammatone filterbank, because it represents a good compromise between accuracy and computational efficiency. The model is based on the assumption that three reservoirs of transmitter substance exist within the hair cell, and that transmitter is released in proportion to the degree of basilar membrane displacement. The amount of transmitter released is equated with the probability of a spike being generated in the associated auditory nerve fiber. The Meddis model replicates many of the characteristics of auditory nerve responses discussed above, including rectification, compression, spontaneous firing, saturation effects and adaptation.

Panel A of Fig. 1.7 shows the response of the Meddis hair cell model to a pure tone stimulus, where the basilar membrane displacement has been provided by a

---

**Figure 1.6** Gammatone filters. (A) Impulse responses for eight gammatone filters, with center frequencies equally spaced between 100 Hz and 4 kHz on the ERB-rate scale of Glasberg and Moore [46]. (B) Frequency responses of the filters shown in panel A.
gammatone filter channel centered on the tone frequency. Panel B of the figure shows the response of a complete cochlear model (gammatone filterbank and Meddis hair cell model) to three periods of a steady-state vowel. Within each frequency channel, the response can be interpreted as the instantaneous firing rate within an auditory nerve fiber. The resulting time-frequency representation is often termed a neural activity pattern or cochleagram. A cochleagram may also refer to a simpler form of cochlear representation: filtering by an auditory filterbank followed by some form of nonlinear rectification.

For visual representation of long signals, the waveform display used in panel B of Fig. 1.7 contains too detailed information. A representation that is more like a conventional spectrogram can be obtained by smoothing the time series associated with each frequency channel, downsampling, and then mapping the resulting values to color or a gray scale. An example of a cochleagram computed in this way is shown in Fig. 1.8, for the utterance "they enjoy it when I audition" spoken by a male talker. The figure was produced by smoothing the output from the Meddis hair cell model with a Hanning window of width 20 ms, and sampling the output at 5 ms intervals. A conventional spectrogram is also shown in the figure, which was also produced using Hanning-windowed frames of width 20 ms and a frame interval of 5 ms. A number of differences between the two representations can be noted. First, the representation of low frequencies is expanded in the cochleagram, because of the ERB-rate spacing of the filter center frequencies. Second, the first formant seen in the spectrogram is resolved into harmonics in the cochleagram, because auditory filters with low center frequencies have a narrow bandwidth. Finally, adaptation in the hair cell model causes acoustic onsets to be emphasized in the cochleagram.

1.3.3 Correlogram

The autocorrelation-based theory of pitch perception proposed by Licklider [77] has formed the basis for many computational models of fundamental frequency (F0) estimation and F0-based sound separation. The first computational implementation of Licklider’s scheme was demonstrated by Richard Lyon, and later published by Lyon [84] and Weintraub [135, 136]. These authors referred to it as an “auto-coincidence” or “coincidence” representation. Subsequently, Slaney and Lyon [119] introduced the term correlogram.

The correlogram is usually computed in the time domain, by autocorrelating the simulated auditory nerve firing activity at the output of each cochlear filter channel:

\[
acf(n, c, \tau) = \sum_{k=0}^{K-1} a(n-k, c)a(n-k-\tau, c)h(k)
\]

Here, \(a(n, c)\) represents the simulated auditory nerve response for frequency channel \(c\) at discrete time \(n\), and \(\tau\) is the time lag. The autocorrelation is computed over a window of length \(K\) samples, which is shaped by a window function \(h\) (which is typically chosen to be a Hanning, exponential, or rectangular shape). For greater efficiency, the correlogram can also be computed in the frequency domain using the...
discrete Fourier transform (DFT) and its inverse transform (IDFT), using the relation

$$acf(x_{\text{nw}}) = \text{IDFT} \{ \text{DFT}(x_{\text{nw}}) \}^p$$  \hspace{1cm} (1.8)

where $x_{\text{nw}}$ is a windowed section of the simulated auditory nerve activity obtained from frequency channel $c$, starting at time $n$. In Eq. 1.8, the use of a parameter $p$ allows for a generalized autocorrelation. Setting $p = 2$ gives a conventional autocorrelation function, but smaller values of $p$ may be advantageous because they yield sharper peaks in the resulting function [124].

Panel A of Fig. 1.9 shows a correlogram for a vowel /æ/ with $F_0 = 100$ Hz, which is derived from the neural activity pattern shown in Fig. 1.7. The figure shows one time frame of the three-dimensional volume defined by $acf(n, c, \tau)$, in which frequency channel and autocorrelation lag are plotted on orthogonal axes. It should be noted that the squaring inherent in the autocorrelation function means that $acf(n, c, \tau)$ requires twice as much dynamic range as $a(n, c)$; in the figure, this has been addressed by normalizing each channel to a maximum value of one.

The correlogram is important as a model of pitch perception because it unifies two schools of pitch theories: the place theories (which emphasize the role of resolved harmonics) and the temporal theories (which emphasize the role of unresolved harmonics [77, 96]. Figure 1.9 demonstrates that the correlogram combines pitch information from both resolved and unresolved harmonic regions. Consider the lowest frequency channels in the figure, which are excited by the $F_0$ of the vowel (100 Hz). They exhibit a peak at the period of the $F_0$ (10 ms), but a peak also occurs at 20 ms because the autocorrelation of a periodic signal is itself a periodic function. Similarly, channels driven by the second harmonic exhibit a peak at the period of 200 Hz (5 ms), but also at 10, 15, 20, 25 ms, and so on. Hence, a peak at 10 ms occurs in all correlogram channels that are excited by a resolved harmonic. At high frequencies, auditory filters are wider and a number of harmonics interact within the same filter. This results in amplitude modulation (“beating”) at the rate of the $F_0$. Hence, the high-frequency channels of the correlogram also exhibit a peak at the period of the $F_0$, and the correlogram for a periodic sound displays a characteristic “spine” that is centered on the fundamental period (and its multiples).

This $F_0$-related structure can be emphasized by pooling the information in the correlogram across frequency. The resulting summary correlogram is given by

$$sacf(n, \tau) = \sum_c acf(n, c, \tau)$$  \hspace{1cm} (1.9)

The summary correlogram is shown below the correlogram in panel A of Fig. 1.9, and exhibits two strong peaks. The peak that occurs at the shortest lag (10 ms) corresponds to the fundamental period. Meddis and Hewitt [89, 90] have demonstrated that the position of peaks in the summary correlogram corresponds closely to perceived pitch. Furthermore, concurrent sources with different $F_0$s give rise to multiple peaks in the summary correlogram, which can be used as the basis for multipitch tracking and $F_0$-based sound separation algorithms (see Chapters 2 and 3).

### 1.3.4 Cross-Correlogram

As with the correlogram, the basic idea underpinning the cross-correlogram was described in the literature some 30 years before a computational implementation was attempted. Specifically, the cross-correlogram is motivated by the work of Jeffress [62], who describes a mechanism by which human listeners might localize sound using interaural time differences (ITD). Jeffress proposes a neural circuit in which firing activity that arises from the same critical band of each ear travels along a pair of delay lines. When the net delay corresponds to the interaural time difference, activity in the two delay lines coincides and is detected by a coincidence unit. This mechanism is equivalent to an interaural cross-correlation of the form

$$ccf(n, c, \tau) = \sum_{k=0}^{M-1} a_l(n-k, c)a_r(n-k-\tau, c) h(k)$$  \hspace{1cm} (1.10)

where $a_l(n, c)$ and $a_r(n, c)$ represent the simulated auditory nerve response at discrete time $n$ and frequency channel $c$, for the left and right ears, respectively. The parameter $\tau$ is the lag, and the cross-correlation is computed over a window $h$ of size $M$ samples (again, a Hanning, exponential, or rectangular window is often
used). Cross-correlogram representations of this form have been used in CASA algorithms by Lyon [83] and many others.

Panel B of Fig. 1.9 shows a cross-correlogram for a steady-state vowel /æ/, which has been presented to a binaural auditory model with an interaural time difference of 0.25 ms. As with the correlogram, the cross-correlogram is actually a three-dimensional volume defined by $ccf(a, c, \tau)$ and the figure shows a slice through this volume taken at a particular time instant. Again, each channel of the cross-correlogram has been normalized to a maximum value of one. The cross-correlogram has a spine centered on the ITD of the stimulus, which is surrounded by “sidelobes” due to harmonic components and filterbank resonances. Changes in the ITD of the stimulus simply cause this pattern to be shifted to the left or the right.

The spinal structure in the cross-correlogram can be emphasized by pooling over frequency, in a manner analogous to that described by Eq. 1.9. The resulting summary cross-correlogram is shown beneath the cross-correlogram in Fig. 1.9, and has a peak at 0.25 ms, which corresponds to the ITD of the stimulus. If the input signal consists of a mixture of sound sources that originate from different locations, then the summary cross-correlogram will generally contain multiple peaks that can be used as the basis for localizing and separating each constituent source (see Chapters 5 and 6).

It should be noted that in the form described above, the cross-correlogram is only sensitive to ITD; interaural intensity differences (IID) do not cause a change in the response pattern. Subsequent studies have introduced IID sensitivity into the dual delay-line model by using an inhibitory input from the contralateral ear [79]. Additionally, various enhancements of the basic cross-correlogram scheme have been proposed that allow it to explain aspects of the precedence effect [80] and binaural signal detection [26]. Finally, the notion of an interaural cross-correlation mechanism is broadly supported by physiological studies, which have revealed systematic arrangement of ITD-sensitive neurons in the auditory midbrain (e.g., [140]).

### 1.3.5 Time-Frequency Masks

Many CASA systems achieve source segregation by computing a *mask* to weight a T-F representation of the acoustic input such as a cochleagram. The mask applies a weight to each T-F unit, such that spectrotemporal regions that are dominated by the target source are emphasized, and regions that are dominated by other sources are suppressed. The values in the mask may be binary or real-valued; in the latter case, the mask value may be interpreted as the ratio of target energy to mixture energy (as in a Wiener filter) or the probability that the T-F unit “belongs” to the target source. A time-frequency weighting of this kind was first employed in the binaural source separation algorithm described by Lyon [83], and has subsequently been adopted by other workers [136, 15, 133, 115].

Perceptually, the use of binary T-F masks is motivated by the phenomenon of masking in auditory perception, in which a sound is rendered inaudible by a louder sound within a critical band (for a review, see [96]). Additionally, different lines of computational consideration have converged on the use of binary masks as described in the following text.

First, Joujine et al. [65] and Roweis [116] have noted that a speech signal is sparsely distributed in a high-resolution T-F representation and, as a result, different speech utterances tend not to overlap in individual T-F units. This observation leads to the property of orthogonality between different speech utterances. In this case, binary masks are sufficient for decomposing sound mixtures into their constituent sources. The orthogonality assumption holds well for mixtures of speech and other sparsely distributed signals (e.g., complex tones), but is not valid for speech babble or other broadband intrusions.

Second, the notion of a time-frequency mask is central to the missing data approach to ASR proposed by Cooke et al. [29]. In this approach, the mask indicates whether each acoustic feature should be regarded as reliable or unreliable evidence of a target speech source, so that it can be treated appropriately during recognition. Furthermore, Cooke et al. introduce the notion of an a priori mask, which is used to assess the upper limit on the performance of a missing data ASR system that uses binary masks. The a priori mask is so called because it is computed using speech and noise signals before mixing. Specifically, Cooke et al. employ a priori masks in which reliable time-frequency regions are assumed to be those for which the mixture energy lies within 3 dB of the energy in the premixing speech signal.

Third, Wang and colleagues [57, 113, 58] have suggested an ideal binary mask as a computational goal of CASA (see Section 1.2.2). Specifically, denote the target energy in a T-F unit as $s(t, f)$ and the interference energy as $n(t, f)$. The ideal binary mask is given by

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

(1.11)

where $t$ and $f$ index the time and frequency dimensions, respectively. The parameter $\theta$ is typically chosen to be 0, corresponding to a 0 dB criterion. Figure 1.10 shows an ideal time-frequency mask for a mixture of speech and a telephone sound, together with cochleagrams for the mixture and clean signals. Figure 1.5, panel E, illustrates the result of applying the ideal mask to the mixture; it is clear from the figure that the masked mixture is much closer to the clean target than the mixture itself.

Support for this notion comes from the observation that target sound reconstructed from an ideal binary mask is of high perceptual quality [115, 113]. The ideal binary mask also provides a very effective front end for automatic speech recognition in noise [29, 113]. Recent experiments with human listeners have found that ideal masking leads to substantial improvements of speech intelligibility; in particular, the ideal binary mask defined where $\theta = -6$ dB appears to be most effective for human speech intelligibility [113, 23, 21].

### 1.3.6 Resynthesis

A number of CASA systems include a resynthesis pathway, which allows an audio waveform to be reconstructed from a group of segments that correspond to an indi-
put from each filter channel is then divided into time frames by windowing with a raised cosine, with a frame size that matches the original decomposition into T-F units. The energy in each T-F unit is then weighted by the corresponding T-F mask value (which may be binary or real-valued between 0 and 1). The weighted responses are then summed across all frequency channels to yield a reconstructed audio waveform. High-quality resynthesis has been obtained in this way for speech and other signal types.

1.4 CASA EVALUATION

Quantitative evaluation of CASA systems is critical for gauging the progress made in the field. Systematic evaluation also serves to guide the development of a model. The evaluation process typically consists of choosing an evaluation criterion or metric, and selecting an evaluation corpus that is representative of the application domain. These two aspects are addressed below.

1.4.1 Evaluation Criteria

A variety of evaluation criteria have been used for CASA systems, which can be broadly divided into four categories: comparisons between segregated target and clean (premixing) target, changes in automatic speech/speaker recognition score, human listening tests, and correspondence with biological data. Each category is described below.

Comparison with Clean Target Signal. This approach assumes that the acoustic mixture to be separated has been generated by the addition of one or more interfering signals to a clean target signal. A ground-truth target signal is therefore available, which can be compared with the segregated target in a number of ways. For example, Cooke [28] assesses the similarity between symbolic time-frequency representations of the clean target (a speech signal) and the segregated target. Brown and Cooke [16] use binary T-F masks to resynthesize two signals, which correspond to the segregated target speech and the residual noise that it contains. These two waveforms are then compared in order to derive a normalized SNR in the range 0 to 1. Wang and Brown [133] use the same strategy, but quantify the performance using a conventional SNR metric that is expressed in decibels. Nakatani et al. [99] use the distortion between the short-term spectra of a segregated source and those of the original source to measure performance for speech separation.

Automatic Recognition Measure. One of the main motivations for speech separation is to improve ASR performance in the presence of acoustic interference. Hence, it is natural to evaluate a CASA system as the front end for an ASR device, so that the recognition score can be compared before and after the target speech is segregated (for example, see [136]). A similar approach can be applied to the evaluation of automatic speaker recognition systems. In the case of automatic music tran-
scription systems, the evaluation typically involves a comparison between the notes in a musical score and the fundamental frequencies detected in the corresponding audio waveform (e.g., [48, 70]).

**Human Listening.** Human listeners can be used to evaluate CASA systems, by conducting formal tests that compare the intelligibility of segregated speech and the unprocessed mixture of speech and noise (e.g., [123]). However, a potential pitfall needs to be considered here: we cannot “turn off” the ASA process of listeners when they are presented with an acoustic stimulus. As a result, we should not expect that a CASA system will yield an intelligibility gain until machine performance exceeds human performance. Hearing-impaired listeners are probably better suited to such an evaluation approach, as it is well known that such listeners have difficulty in recognizing speech in a noisy environment [95]. Certainly, if the intended application of the CASA system is to improve the hearing of impaired listeners, or that of normal listeners in very noisy environments, then this evaluation methodology is an appropriate choice. A different approach to CASA evaluation using human listeners has also been described by Ellis [39]. He asked listeners to score the resemblance between segregated sounds and the corresponding component sounds in an acoustic mixture.

**Correspondence with Biological Data.** If biological plausibility is an objective, then the evaluation criterion is usually how well the computer model of ASA accounts for known perceptual or neurobiological data. Models that employ this criterion include those of Wang [131], who simulated several qualitative ASA phenomena on the basis of a neural oscillator network (see [101] for a recent extension that produces quantitative results); McCabe and Denham [86] who proposed a neural network to model auditory streaming data; and Wrigley and Brown [138], who put forward a neural oscillator model to quantitatively simulate auditory attention data.

It has often been pointed out that there is no consensus on how to evaluate a CASA system [114, 40, 132]. This partly reflects the fact that idiosyncratic evaluation criteria are largely unavoidable in a maturing field. Also, the diverse range of CASA applications (see Section 1.2.3) necessitates different performance measures (for example, human intelligibility testing would be appropriate if the intended application were a hearing aid, but not if the application were ASR). These considerations notwithstanding, it is important to make the computational objective explicit and evaluate accordingly. For example, if the ideal binary mask is the goal, then one can measure the SNR using the signal resynthesized from the ideal mask [58]. Also, wherever possible, one should make an effort to use common evaluation tasks [40].

### 1.4.2 Corpora

A diverse range of corpora have been used for evaluating CASA systems. For voiced speech segregation, Cooke [28] compiled a list of 10 voiced utterances—five sentences spoken by two male speakers—and 10 noise intrusions. The noise samples vary in their bandwidth and temporal structure, and include a recording from a busy teaching laboratory (which is similar to “cocktail party” noise), rock music, narrowband and broadband sounds, and other speech utterances. Combining each voiced utterance with each noise intrusion gives 100 mixtures, which have been commonly employed in CASA studies and facilitated quantitative comparison of different approaches (see Chapter 3). For general speech segregation, one can retain Cooke’s noise intrusions while replacing the voiced utterances with more naturalistic speech material, such as those from the TIMIT corpus [44] or the TIDigits corpus [74].

Other corpora of noise intrusions include the NOISEx database [128], which contains various artificial recordings such as factory noise and tank noise. Hu [56] collected 100 environmental sounds such as rain, wind, and crowd sounds, and used them for training and testing during the evaluation of an unvoiced speech separation system.

For evaluating ASR in noise, AURORA has become a standard series of corpora [107]. Presently there are three AURORA corpora of increasing complexity, in which the target speech is either digits or the 5000 word Wall Street Journal corpus [106]. The background noise (e.g., car noise) is either added digitally or recorded together with target speech, and a range of different SNR conditions are available. To facilitate comparison among front-end processing methods, standard back-end speech recognizers are also provided for reference.

Many tests (and corpora) have been devised to evaluate speech intelligibility in noise by both normal-hearing and hearing-impaired listeners. For example, the speech perception in noise test (SPIN) [67, 9] evaluates word recognition in context. Each SPIN list includes 50 sentences with 25 high-context sentences and 25 low-context sentences, and the listener’s task is to repeat the last word of each sentence that is mixed with multitalker babble. Another example is the hearing in noise test (HINT) corpus [100]. The HINT corpus consists of 25 phonemically balanced lists, each containing 10 short sentences, mixed with speech-shaped noise. The sentences in HINT were adapted for use in American English from the BKB (Bamford–Kowal–Bench) corpus that contains short, semantically predictable sentences spoken by British speakers [5].

Relatively few corpora have been recorded in real acoustic environments with CASA evaluation in mind. A notable exception is the ShATR corpus [69], which was designed specifically for evaluating CASA systems. Recordings were made via multiple microphones, including close-talking microphones for each participant, an omnidirectional microphone, and a binaural dummy head. Related is the ICSI meeting corpus [61], which records multiparty meetings via both near-field and far-field microphones. In significant portions of the ICSI corpus, speech is mixed with other speech and nonspeech noise. Evaluation is facilitated by the availability of Reference signals recorded by near-field microphones, but even these can contain some degree of interference from other speakers (“crosstalk”) [139].

The above corpora are specifically targeted at speech processing. Corpora exist for other sounds, including bird calls and music. In the latter case, the RWC music database [49] is a good resource for CASA systems concerned with music process-
ing, since it contains audio files and corresponding ground-truth data in the form of MIDI transcriptions and lyrics. Other useful corpora include collections of head-related transfer functions (HRTFs) recorded from dummy heads, which can be used to spatialize sound sources for evaluating binaural CASA systems. The website www.casabook.org provides a comprehensive list of these resources in addition to the aforementioned corpora.

Let us end this section by asking the question, What evaluation results will lead us to declare that the CASA problem is solved? The answer to this question obviously depends upon what is expected from CASA. If the purpose is an application domain, such as noise-robust ASR, then success can be more readily verified. Indeed, in certain restricted conditions, machine performance has already matched or surpassed that of human listeners. However, we favor a general criterion: we consider the problem solved when a CASA system achieves human ASA performance in all listening situations. It is the versatility of the human auditory system that, we believe, CASA should aim for.

1.5 OTHER SOUND SEPARATION APPROACHES

The sound separation problem has been investigated in audio signal processing for many years, and is usually motivated by specific engineering problems (e.g., robust ASR in the cockpit of an aircraft, where the statistics of the background noise are known). Two main approaches have been proposed: beamforming and independent component analysis. Each is explained and contrasted with CASA below. We also consider speech enhancement, which is related to source separation but has a narrower goal—to enhance target speech that is contaminated with noise.

Beamforming achieves sound separation by using the principle of spatial filtering ([127, 72, 12]; see also Chapter 6). Spatial filtering aims to boost the signal coming from a specific direction by appropriate configuration of a microphone array, and in doing so it attenuates interfering signals from other directions. The simplest is a delay-and-sum beamformer, which adds multiple microphone signals from the target direction in phase and uses phase differences to attenuate signals from other directions. The amount of noise attenuation increases as the number of microphones and the array length increase. Adaptive beamforming attempts to cancel noise sources via a weight adaptation or training process. In general, an adaptive beamformer with L microphones can remove only L − 1 different noise sources. To overcome this limitation, subband versions of adaptive beamforming have been developed, which allow cancellation of more noise sources whose spectra do not substantially overlap. With a properly configured array, spatial filtering can produce high-fidelity separation. Another advantage of spatial filtering is its ability to attenuate reverberation, because the target signal has a specific direction, whereas reverberant signals arrive from diffuse directions (see Chapter 7). The main limitation of beamforming is configuration stationarity [132]; it is difficult to separate a target that moves around or switches between different sound sources, both of which occur in “cocktail party” situations. Also, no separation is possible when multiple sounds come from directions that are similar or near to each other.

An approach related to beamforming is blind source separation using independent component analysis (ICA) [66, 4, 60], which combines adaptive filtering and machine learning techniques. Like beamforming, mixture signals are modeled in the standard form as a linear superposition of source signals. In other words, a mixing model of the form x(t) = As(t) is assumed, where s(t) is a vector of unknown source signals, A is a mixing matrix, and x(t) is a vector of the mixed signals recorded by multiple microphones. In ICA, the main assumption is that sound sources are statistically independent. The separation problem is formulated as that of estimating a demixing matrix (i.e., the inverse of A). To make this formulation work requires a number of assumptions about the mixing process and the number of microphones [125]. ICA gives impressive separation results when the assumptions are satisfied, but its scope is limited as a result. For example, there must be at least as many microphones as the number of sources (although efforts have been made to relax this constraint [73, 137]). A more fundamental limitation is that the mixing matrix A needs to be stationary for a period of time in order to allow estimation of a large number of parameters. This assumption—akin to the configuration stationarity assumption in adaptive beamforming—is difficult to satisfy in situations in which speakers turn their heads or move around. Like spatial filtering, to be separable, source signals must originate from different spatial directions.

The goal of speech enhancement is to enhance target speech in the presence of background noise [78, 6]. Generally intended to apply to single-channel audio, this approach is based on statistical analysis of speech and noise, followed by estimation of clean speech from noisy speech. Many speech enhancement techniques have been proposed, including spectral subtraction and mean-square error estimation. The widely used method of spectral subtraction subtracts the power spectral density of the estimated background from that of the mixture [11]. The mean-square error estimator, proposed by Ephraim and Malah [41], models speech and noise spectra as statistically independent Gaussian random variables, and optimally estimates clean speech according to the minimum mean-square error criterion. To apply speech enhancement requires a good estimate of the interferer, which is difficult to obtain unless the interferer is stationary. To relax the stationarity assumption, algorithms typically involve detection of speech silence and subsequent noise estimation within speech gaps. Generally speaking, speech enhancement views a mixture as speech plus noise. This perspective is fundamentally narrower than the ASA perspective, which views a mixture as an auditory scene in which a variety of sounds can appear or disappear unpredictably. Speech enhancement cannot deal with such situations.

In visual processing, there is a subtle distinction between the terms “computer vision” and “computational vision.” The former is more concerned with image processing and is more application-driven, whereas the latter is more concerned with modeling human vision as epitomized by Marr’s research [85]. The distinction is analogous to that between other sound separation approaches and CASA.
1.6 A BRIEF HISTORY OF CASA (PRIOR TO 2000)

The following is not intended to be an exhaustive review of CASA development; rather, we focus on a number of key studies and emphasize their historical relationship and the main innovations introduced by each. Our account covers the period before the year 2000, as it takes time for a study to show its historical significance; also the subsequent chapters highlight many recent developments.

1.6.1 Monaural CASA Systems

The problem of monaural sound separation was first brought to a wide audience by Parsons [103], who describes a frequency-domain approach for the segregation of concurrent voices recorded by a single audio channel. Although Parsons makes reference to the “cocktail party effect,” his approach is founded upon conventional signal-processing techniques rather than auditory modeling, and is principally motivated by the problem of crosstalk on communication channels. This system separates two talkers by selecting the harmonics belonging to each, and is therefore limited in its scope because each utterance must be continuously voiced. However, Parson’s study established some directions for research that have proved influential, namely multipitch tracking, F0-based sound segregation, and the use of temporal continuity constraints. His work also represents an important conceptual advance, since its goal was to separate both voices from an acoustic mixture. It is therefore distinct from the preceding literature on speech enhancement, which aimed only to enhance a target voice (for a collection of early speech enhancement papers, see [78]).

A system for segregating the voices of two talkers described by Weintraub [136] may be regarded as the first monaural CASA study of any sophistication. In particular, it stands as the first approach that was grounded in a theory of auditory processing, rather than purely engineering concerns. Weintraub considers a harder problem than Parsons, in that each of the two speech signals can be voiced, unvoiced, or silent at any given time. To address this, Weintraub uses a state-based tracking procedure in which the number of voices present and their characteristics are represented by a Markov model. This approach enables sequential and simultaneous grouping to be represented within a common framework, and represents a significant advance over the simple continuity constraints used by Parsons. Voiced speech is separated on the basis of F0, but a joint time-frequency approach is employed—the correlogram—rather than a frequency-domain approach. Weintraub also develops a framework for evaluating his system. A waveform for each separated voice is resynthesized by refiltering the input waveform, weighted by spectral estimates from the Markov model. This waveform is then used as the input to an ASR system, allowing a comparison of speech recognition accuracy before and after processing by the system.

Weintraub’s model has several limitations. For example, the state-based tracking requires that the two voices have a different average pitch range (a male and female talker were used), and therefore avoids the general problem of sequential grouping. Although unvoiced speech is considered, the model actually does little to segregate it. Also, the results of the evaluation are inconclusive; in some conditions, processing by the system actually causes a deterioration in ASR performance. Weintraub acknowledges that the link between source separation and ASR is a weakness of his system, but makes a number of suggestions as to how this could be improved. In particular, his suggestion that ASR systems should treat silent and masked intervals differently during decoding ([136], p. 145) comes very close to the notion of “missing data” speech recognition adopted by subsequent workers (see Chapter 9).

The study by Cooke [28] appeared shortly after the publication of Bregman’s book [13], and is strongly influenced by his account. Additionally, Cooke was one of the first to note the relationship between Bregman’s ASA theory and Marr’s computational framework of vision [85], and to argue for a greater emphasis on the representation of the acoustic signal than his predecessors. More specifically, he proposes a time-frequency representation called synchrony strands, which is derived by applying local similarity and continuity constraints to the output of a cochlear model. Frequency, amplitude, and amplitude modulation rate are computed at each point along the strands, forming a rich representation that is amenable to search. This approach can potentially separate two voiced utterances with crossing F0s, a challenging problem for earlier systems. Cooke’s system achieves separation of the two sources by grouping their harmonics in the time-frequency plane, so no explicit F0 tracking is necessary. Unlike previous approaches, Cooke’s system does not require the number of sound sources to be specified a priori; rather, grouping continues until all the synchrony strands are accounted for, and the number of sources found is therefore an emergent property of the grouping process.

It should be noted that Cooke’s system is not able to sequentially group acoustic events from the same source that are separated by silent intervals. It therefore does not address the problem of allocating a sequence of voiced and unvoiced speech sounds to the same source. Accordingly, Cooke evaluates his system using a corpus of acoustic mixtures in which the target sound is continuously voiced speech. His evaluation scheme is based on a comparison of the synchrony strand representations of two signals before they are mixed and after they have been mixed and grouped by his system. A criticism of this evaluation approach is that it is tightly bound to the synchrony strand representation; it is therefore difficult to compare the performance of Cooke’s system with other CASA approaches that use different representations.

Mellinger [92] completed his study in the same year as Cooke, and also cites Bregman and Marr as theoretical influences. However, Mellinger’s study differs in a number of respects from Cooke’s, not least because his system is specialized for grouping the resolved harmonics of pitched musical instruments. Additionally, Mellinger’s system does not use a discrete representation such as synchrony strands; rather, the output of a cochlear model is processed directly to form “feature maps” that encode acoustic onsets and frequency modulation. Harmonic components are grouped by considering the affinity between them, which is initially determined by their onset synchrony and subsequently by the coherence of their FM. Mellinger’s system works in an online manner, so that signal representations are
computed and grouped as time progresses. This is distinct from the “batch processing” of Cooke’s system, in which the synchrony strand representation is computed first for the whole input, and then grouped. Online processing is clearly more suited to real-world applications.

Mellinger notes that his system “fails to work in some musically significant situations” (p. 183). This can be attributed to a number of factors. First, despite the important role of harmonic relations in music, his system has no way of grouping harmonics unless they share a common onset time and exhibit the same FM. Second, in his system the tracking of a harmonic is triggered by the detection of an acoustic onset. This approach is somewhat fragile; failure to detect an onset means that an entire acoustic event is missed by his tracking procedure.

The study of Brown [15] (see also [16]) combines a number of ideas from previous systems. Like Mellinger, his system employs maps of feature detectors that are justified on physiological grounds. In Brown’s system, these maps extract onset, offset, periodicity, and frequency transition information from the output of a cochlear model. Periodicity information is derived from the correlogram, and is used for detecting resonances due to harmonics and formants. However, in Brown’s system these intermediate representations are used to construct discrete time-frequency elements (or segments), which are similar in concept to Cooke’s synchrony strands. Recall that in Cooke’s system, specific grouping mechanisms are required to combine resolved and unresolved harmonics of the same voice. This step is avoided in Brown’s study by deriving a pitch contour for each segment using “local” correlograms. This mechanism works in the same way for both resolved and unresolved harmonics, thus allowing them to be considered by a single search algorithm. Brown’s system therefore represents an interesting fusion of ideas from Weintraub and Cooke’s approaches; it is based on the correlogram, but performs grouping on the cochleagram using pitch contours that are computed for individual segments. Again, this means that there is no requirement for a priori knowledge of the number of sound sources that are present in the input.

Brown evaluates his system on the same corpus of voiced speech and noise intrusions used by Cooke, although the evaluation approach is quite different. It is assumed that the energy at each point in the time-frequency plane is allocated to a single source (the principle of “exclusive allocation” [13]) and hence the result of grouping is represented as a binary time-frequency mask. A signal for a target source is resynthesized by refiltering the input and weighting it by the time-frequency mask, in a similar approach to that described by Weintraub (see Section 1.3.6). However, Brown extends this approach by resynthesizing the original sources (before mixing) from the mask, so that the proportion of signal and noise energy in the separated target signal can be determined. This allows Brown to express the performance of his system in terms of an intuitive signal-to-noise ratio (SNR) metric.

Subsequently, innovations in CASA tended to focus on the system architecture. A novel contribution in this regard is the residue-driven architecture introduced by Nakatani et al. [99]. Critical of the “batch processing” approach of Cooke and Brown, these authors proposed an online scheme within a multi-agent paradigm. In their system, grouping processes are formulated as (largely independent) agents that interact to explain the auditory scene. Agents are of three types: event detectors, tracer generators, and tracers. Event detectors subtract the predicted input from the actual input, creating a residue. Tracer generators evaluate the residue and determine whether there is sufficient evidence to spawn a new tracer. Tracers group the features corresponding to a single stream, and also predict what the next input will be. The predicted input is then communicated to the event detector, thus completing the processing cycle.

Nakatani et al. describe agents for tracing static background noise and harmonics. For example, noisy speech might be accounted for by a noise tracer and a harmonic tracer. However, if a substantial harmonic residue is found during processing of the input, this suggests that another source has appeared and a new harmonic tracer is generated to track it. The authors evaluate the residue-driven system by comparing the F0 contours derived by harmonic tracers with ground-truth F0 contours, and by measuring the spectral distortion between the original and recovered sources. Additionally, they demonstrate the flexibility of the agent-based architecture by showing the variation in its performance when different types of agents are added and removed (e.g., its performance with and without noise tracers).

The use of a feedback loop in the residue-driven architecture was a development that subsequently motivated Ellis [39] to propose a prediction-driven approach to CASA. Ellis characterizes previous CASA systems as “data-driven,” and argues that certain phenomena (such as the interpretation of masked signals in the auditory continuity effect) cannot be modeled without top-down information. He therefore proposed a system that is based on an internal world model, in which the auditory scene is explained in terms of a hierarchy of sound elements. The world model is constantly reconciled with the acoustic input, according to hypotheses of the sound sources present.

The prediction-driven architecture is implemented as a blackboard system, in which independent knowledge sources communicate via a shared database of hypotheses (the “blackboard”). In this respect, the prediction-driven approach to CASA may be regarded as a descendent of earlier blackboard-based signal understanding systems such as IPUS [75]. Ellis notes that the blackboard may have a conceptual advantage over the residue-driven approach of Nakatani et al. In the residue-driven approach, the detection, prediction and search for certain kinds of signals (such as harmonically related components) are tightly bound by relatively constrained interactions between specialized agents. Thus, it is not immediately clear how different properties of the same signal (such as its harmonic relations and spatial location) should be combined. In contrast, the blackboard architecture provides a relatively unconstrained framework in which various knowledge sources can influence alternative hypotheses about the input.

Ellis made a number of other contributions in this work. First, his CASA system uses a broader signal model than previous approaches, and includes representations of periodic elements derived from the correlogram (“wells”), aperiodic elements (“noise clouds”), and transients. Second, his system allows the energy in a specific time-frequency region to be shared between sources, since harmonic sounds are
their system identifies low-frequency envelope modulations that are typical of speech signals. The resulting AM spectrum is then weighted by a function derived from a localization system, so that speech energy that arises from a particular direction is passed, and other energy is suppressed.

There have also been attempts to incorporate binaural cues into more sophisticated CASA architectures. In particular, Nakatani et al. [98] have incorporated binaural cues into their residue-driven architecture. Their motivation for doing so is to address the poor performance of their system using harmonic tracking only, when the F0s of two sound sources are very close. Their binaural system initially uses harmonicity to separate a sound source, until a stable estimate of its location has been determined over a period of time using intensity and phase differences received at two microphones. Information about the source direction is then used to predict the next input, so that harmonic fragments are allocated to the correct location even if their F0s are close together and harmonic tracking fails.

### 1.6.3 Neural CASA Models

Little in Bregman’s account [13] alludes to the neural mechanisms that underlie ASA. Similarly, most workers developing CASA systems have made clear that their systems are functional models rather than physiologically plausible ones. What, then, is the neural basis for ASA?

A partial answer to this question is suggested by the studies of Beauvois and Meddis [3], McCabe and Denham [86], and Grossberg [50], who describe neural models that exhibit the auditory streaming phenomenon. However, a more general account of the neural mechanisms underlying ASA is suggested by von der Malsburg and Schneider [129], who propose a neural network model of auditory source separation in which the allocation of features to streams is determined by neurons with an oscillatory response. Specifically, oscillators that represent features of the same stream are synchronized (phase locked with zero phase lag), and are desynchronized from oscillators that represent different streams.

A similar approach was subsequently adopted by Wang [130, 131], who models auditory stream segregation using a network of locally connected excitatory units with a global inhibitor. However, Wang addresses a fundamental problem apparent in von der Malsburg and Schneider’s approach, namely that their oscillator network is dimensionless and fully connected; it therefore indiscriminately connects oscillators that are activated simultaneously by different acoustic sources. Instead, Wang hypothesizes the use of delay lines to construct a neural time axis, such that the auditory scene is represented within a time-frequency grid of oscillators.

These two studies establish the principle of oscillatory correlation as a neural basis for ASA, but both required input in the form of symbolic patterns. Subsequently, Brown and Cooke developed two systems that paired a cochlear model with a network of chaotic oscillators, and processed sampled audio signals. The first modeled the grouping of acoustic components by harmonic relations and onset synchrony [18], and the second modeled a variety of two-tone streaming phenomena [19].
Following this, Wang and Brown [20] combined their approaches in a model of double vowel separation that took acoustic signals as input, and was based on the LEQON (locally excitatory, globally inhibitory) network of Terman and Wang [134]. This led to the development of a complete CASA system within the neural oscillator framework [133]. In this system, the two conceptual stages of ASA—segmentation and grouping—are represented in two layers of a LEQON network. Each layer is a two-dimensional time-frequency grid of oscillators. In the first layer, synchronized blocks of oscillators (segments) are formed, which correspond to connected regions of energy in the time-frequency plane (such as harmonics and formants). Different segments are desynchronized from one another. In the second (grouping) layer, connections are formed between segments if they are compatible with a fundamental frequency estimate obtained from a correlogram.

Wang and Brown’s system is of interest because of its neurobiological basis. In terms of system architecture, one potential advantage of the neural oscillator approach is its parallel and distributed nature, which should make it more suitable for direct hardware implementation than other CASA architectures. However, complex oscillatory dynamics makes it difficult to incorporate more grouping principles.

1.7 CONCLUSIONS

Systems for CASA vary in their architecture, biological motivation, and dependence upon top-down and bottom-up grouping processes. All, however, have the same overriding goal—to computationally extract descriptions of individual sound sources from one or two recordings of an acoustic scene.

CASA is a relatively young discipline, and although much progress has been made in recent years, we are still far from achieving machine performance that is comparable with that of human listeners. The above review has identified a number of aspects of CASA systems (multipitch tracking, feature-based processing, binaural source localization and grouping, model-based segregation, and neural/perceptual modeling) which are key areas for further research. Robustness to reverberation is also an important goal, as is the application of CASA techniques to automatic speech recognition and music processing. The following chapters discuss these issues in more detail.

ACKNOWLEDGMENTS

Thanks to Guoning Hu and Soundararajan Srinivasan for their assistance with the preparation of this chapter, and to Albert Bregman and Stuart Wrigley for commenting on earlier drafts. Wang’s research was supported in part by an AFOSR grant (FA9550-04-01-0117) and an AFRL grant (FA8750-04-1-0093). Brown’s research was supported by EPSRC grant GR/R47400/01.

REFERENCES


57. G. Hu and D. L. Wang. Speech segregation based on pitch tracking and amplitude


2.1 INTRODUCTION

This chapter is about the estimation of multiple fundamental frequencies (F0) from a waveform, such as the compound sound of several people speaking at the same time, or several musical instruments playing together. That information may be needed to transcribe a musical score, to extract intonation patterns for speech recognition, or as an ingredient for computational auditory scene analysis. The task of estimating the single F0 of an isolated voice has motivated a surprising amount of effort over the years [45]. Work on the harder task of estimating multiple F0s is now gaining momentum, fueled by progress in signal processing techniques on the one hand, and new applications such as interactive processing or indexing of music, multimedia, and speech on the other.

A multiple F0 estimation method is typically assembled from two elements: a single-voice F0 estimator, and a voice-segregation scheme. Here “voice” is used in a wide sense to designate the periodic signal produced by a source (human voice, instrument sound, etc.). Some space is accordingly devoted to the topic of single voice F0 estimation, but the reader should refer to the excellent treatise of Hess [45] for more details. Segregation techniques too are evoked, but the reader should follow pointers to other chapters of this book wherever possible.

A sound with a periodic waveform evokes a pitch that varies with F0, the inverse of the period [87]. The pitch may be salient and musical as long as the F0 is within about 30 Hz to 5 kHz [92, 105]. Sounds with the same period evoke the same pitch despite their diverse timbres, so pitch can be understood as an equivalence class. The auditory system is able to extract the period despite very different waveforms or spectra of sounds at the ears. Explanations of how this is done have been elaborated since antiquity [27]. Modern theories can be classified into two families: pattern matching and autocorrelation [27]. These theories are a source of inspiration for the development of F0 estimation methods that likewise can be organized according to a small number of basic principles, as we shall see in Section 2.3. Quite good solutions now exist for the task of single F0 estimation [45, 31].