Chapter 2

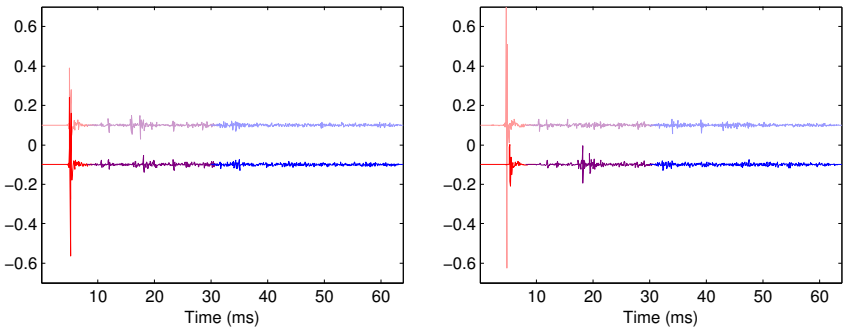
Literature review

This chapter reviews some of the research relevant to source separation and localization by humans and machines. While the literature on these topics is extensive, effort has been made to limit the discussion to areas that are most relevant to the later chapters of this work, namely localization and separation of speech in the presence of other speakers and under realistic spatial listening conditions, including reverberation. In some areas, these specific topics have not been studied in much depth, and we must fall back on studies that use noise maskers or tonal targets instead of speech, headphone listening instead of free-field, or anechoic environments instead of reverberant. There are many excellent reviews of each of these topics and each section begins with references to relevant reviews. Section 2.1 begins the chapter with an overview of the binaural room impulse response, broken down into direct-path, early echo, and late reverberation portions. It describes the information carried in these parts, and their effect on the acoustics and psychoacoustics of binaural listening.

2.1 Anatomy of a binaural room impulse response

The binaural room impulse response (BRIR) is the transfer function from a source to both ears of a human listener (or dummy head). It is influenced by and contains information about the room, the head, the torso and shoulders, and the position and orientation of these things relative to one another. It does not contain information about any particular sound source, only the channel that would convey the sound from that source were it located in that particular configuration. Any signal that travels through this channel will be convolved with this impulse response.

The BRIR is made up of three main components, the direct-path, early echoes, and late reverberation, which have both acoustic and psychoacoustic relevance. See figure 2.1 for an example BRIR with these three components highlighted. Separate versions of each of these components can be simulated from a pre-recorded impulse response by setting various parts of the impulse response to 0.



(a) Source at 0° (b) Source at 75°

Figure 2.1: The first 64 ms of two example binaural room impulse responses (BRIRs). The direct path is red, early echoes (here defined as all energy between 10 and 32 ms) are purple, and late reverberation is blue. The left ear is the lighter lines, the right ear the darker lines. A DC offset is introduced to aid in distinguishing between the two ears.

Mathematically, the binaural room impulse response, $h_\ell(t)$ and $h_r(t)$, can be decomposed into the sum of these three components

$$h_\ell(t) = h^\ell_{dp}(t) + h^\ell_{ee}(t) + h^\ell_{rev}(t) \qquad h_r(t) = h^r_{dp}(t) + h^r_{ee}(t) + h^r_{rev}(t) \tag{2.1}$$

where each h_{dp} is the direct path impulse response for a particular ear, h_{ee} is the early echo impulse response, and h_{rev} is the late reverberation impulse response. The convention used here is to distinguish these components by stipulating that at most one of them is non-zero at a particular time, t, for each ear. In particular, h_{dp} is non-zero first, followed by h_{ee}, and then h_{rev}. Because of the linearity of convolution,

$$s(t) * h_\ell(t) = s(t) * h^\ell_{dp}(t) + s(t) * h^\ell_{ee}(t) + s(t) * h^\ell_{rev}(t). \tag{2.2}$$

Thus, with access to a recording of a BRIR, the direct-path, early echoes, and late reverberation can be individually included, omitted, or manipulated.

The next sections discuss the properties of these three parts of a BRIR. Table 2.1 summarizes these sections and figure 2.2 shows an example sound convolved with these three parts.

2.1.1 Direct-path

The direct-path sound is the sound coming directly from the source. It travels the shortest distance from the source to the listener, and so it is the first to arrive. Its energy falls off as r^2, where r is the distance between the source and listener. In an anechoic chamber, this is the only path that sound takes from source to listener. The direct-path sound contains accurate information about the location of the source, because it arrives from the true direction of the source. Psychoacoustically, this is the component of the impulse response

Table 2.1: A comparison of some of the characteristics of the three components of a binaural room impulse response.

	Direct-path	Early echoes	Late reverberation
Relative timing	0 ms	1–30 ms	$>$ 30 ms
Information about	source location	room geometry	room size, materials
Change with distance	r^{-2}	constant?	constant
Change with motion	moderate	slow	details change rapidly
Effect on intelligibility	improve	improve	diminish
Interaural parameters	main trend	perturb mean	increase variance

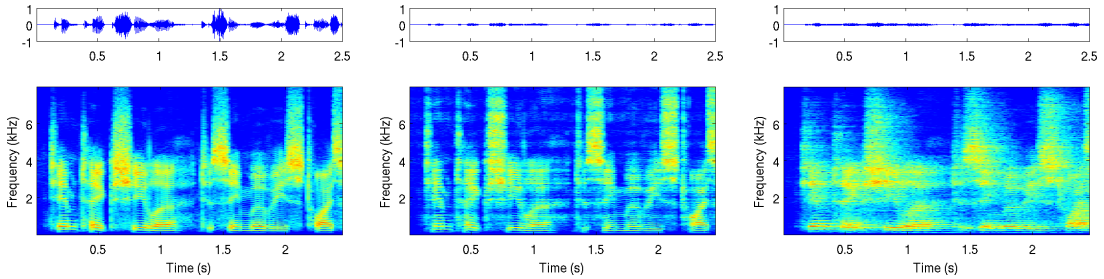
that is used in localizing sources (Litovsky et al., 1999). Because the direct-path sound depends on the relative location of source and listener, it changes with the motion of either one. The source-to-listener distance can be quite small, so it can change with relatively small motions of the source or listener.

The direct-path signal is colored by the head-related transfer function (HRTF), which is the portion of the transfer function between source and listener that depends on the listener's body, including the head, outer ears (pinnae), shoulders, and torso (Algazi et al., 2001a). It depends mainly on the relative azimuth and elevation of the source and listener, varying little with distances greater than about 1 m. It thus contains information useful for localization in both azimuth and elevation.

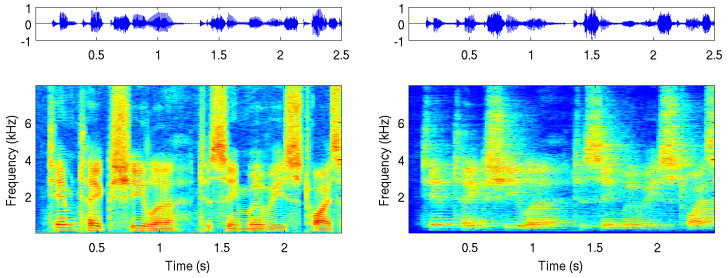
2.1.2 Early echoes

Early echoes arrive immediately after the direct-path sound. Acoustically, these are any echoes that can be considered "specular", i.e. that could be considered distinct sources on their own. Specular reflections can be distinguished by their high coherence (cf. section 4.1.3) at the two ears (Gardner, 1992) . Psychoacoustically, they arrive soon enough after the direct-path sound to be grouped with it into a single percept (Watkins and Holt, 2000). Typically, this means they arrive within 1 to 30 ms of the direct-path (Litovsky et al., 1999), although the upper limit, known in psychoacoustics as the "echo threshold" changes dramatically with the impulse responses used (Ebata et al., 1968) and with experimental conditions (Hidaka et al., 2007). These echoes carry information about the geometry of a space, including such information as its volume and the number and orientation of walls (Benade, 1985). Early echoes also improve the intelligibility of speech by increasing the amount of speech energy arriving at the listener (Aigner and Strutt, 1935; Haas, 1951; Bradley et al., 2003; Watkins and Holt, 2000).

Specular reflections can be simulated with the so-called image model (Allen and Berkley, 1979), in which walls are treated as acoustic mirrors. Because of the law of reflection, a reflection creates a virtual replica of the original source that acts like it is located behind the wall symmetrically. In the simplest case, for a room that is shoe-box shaped, i.e. with six walls that are all either parallel or perpendicular to one another, these virtual sources are arranged in a three-dimensional grid and computation of the impulse response is relatively



(a) Direct-path only (b) Early echoes only (c) Late reverberation only



(a) Original (b) Combined

Figure 2.2: An example utterance convolved with different parts of a binaural room impulse response. Shows the original signal, the signal convolved with the direct-path, early echoes, and late reverberation portions of the impulse response, and the full reverberated signal, in the time and spectrogram domains. Only the left ear signals are shown.

straightforward. For rooms with other geometries, ray tracing can be used to calculate an impulse response using this model (Krokstad et al., 1968; Schroeder, 1970).

Early echoes depend on the geometry of a space, source, and listener. They are generally less sensitive to listener or source movement than the direct-path sound because of the larger source-to-listener distance. The virtual sound sources are farther away than the actual sound source, and so the same absolute motion is smaller relative to these larger source-to-listener distances. In addition, the energy of the early echoes depends less on the source-to-listener distance than that of the direct path because moving farther from one virtual source brings a listener closer to others.

2.1.3 Late reverberation

Late reverberation arrives after the early echoes. Acoustically, it is made up of a very large number of higher order reflections and scattering off of walls and objects. It behaves somewhat like a diffuse source (Morrow, 1969; Izumi et al., 2007), generally does not contain specular reflections, and is incoherent at the two ears (Gardner, 1992) . While individual reflections cannot be discerned, it is useful for its bulk properties, which include

a characterization of the size of the room (Cabrera et al., 2005) and the materials that a room's walls are made of (Huopaniemi et al., 1997). This is because the frequency-dependent reflectance of the walls is amplified many times after many reflections off of surfaces in a room.

Psychoacoustically, late reverberation is the portion of the impulse response after the echo threshold. This could include specular reflections (echoes), as from a far away wall, but typically indoors it includes more diffuse reverberant energy. It actively detracts from speech intelligibility (Lochner and Burger, 1964), especially for non-native speakers (Takata and Nábělek, 1990) and the hearing impaired (Nábělek and Mason, 1981).

The amount of energy in the late reverberation is typically spread throughout a room rather uniformly, and so is the same for a listener in any location in a room. While the direct-path energy falls off as r^2 with distance, the late reverberation energy stays relatively constant. Humans use this ratio, quantized coarsely (Larsen et al., 2008), to distinguish near from far sources (Von Békésy, 1960, pp. 301–313). The so-called critical radius is the distance from a source at which the direct-path and reverberant energies are equal. A related psychoacoustic quantity is the reverberation radius, which is the distance inside of which intelligibility increases because of direct-path energy and beyond which intelligibility remains constant (Libbey and Rogers, 2004; Peutz, 1971). This distance is typically 3–3.5 times the critical radius.

Late reverberation changes very rapidly with the motion of either the source or the listener, as it is made up of a combination of a very large number of reflections and their interaction is very dependent on the phase of those additions. It is also very sensitive to air currents and temperature gradients for the same reason (Omura et al., 1999). While this instability in the fine detail makes it difficult to deconvolve the late reverberation from a signal (Radlovic et al., 2000), it does not change the "sound" of the reverberation to listeners.

2.1.4 Summary

A recorded binaural room impulse response may be decomposed into three distinct components. The direct-path component carries information about the location of sources, and section 2.3 discusses the ways in which humans take advantage of this information. It also aids intelligibility, as do early echoes, while late reverberation detracts from intelligibility. These properties will be discussed further in section 2.3. This decomposition of the BRIR forms the basis of our direct-path, early echo, reverberation, target, masker (DERTM) evaluation metric and is important to the analysis of the automatic speech recognition results in chapter 6.

2.2 Human separation performance

Humans are quite good at listening to a conversational partner in a wide variety of conditions. This section describe some experiments that have measured the ability of humans to understand speech in adverse conditions with the goal of later comparing these results against the performance of our source separation systems. There are a number of

very good review articles on this topic and the interested reader is referred to (Bronkhorst, 2000; Assmann and Summerfield, 2004; Brown and Palomäki, 2006; Darwin, 2007).

The motivating example of the robustness of human hearing is the cocktail party, in which it is possible to carry out a conversation in a room full of similar conversations taking place nearby. Plomp (1977) calculated, after many approximations, that for the signal-to-noise ratio to be greater than 0 dB in such a situation, the speaker and listener must stand no more than 0.7 m apart (Darwin, 2007). Furthermore, without reverberation the speech energy reaching the listener is a mere 2 dB above the threshold required for 95% speech intelligibility, and reverberation can easily remove this advantage (Brown and Palomäki, 2006).

This work exclusively studies binaural means of localizing and separating sound sources. This choice was made because of the interesting property of human hearing that binaural hearing provides an advantage over monaural hearing in many tasks (Cherry, 1953). The simplest examples of this are experiments involving headphone presentation of speech and noise combined with different polarities in time (Licklider, 1948). In this work, a speech target, $s(t)$, and noise masker, $n(t)$, were presented to listeners over headphones at varying relative levels. When the same signal, $n(t) + s(t)$, was presented to both ears of the listener, the speech-to-noise ratio for 50% intelligibility (known as the speech reception threshold, or SRT) was approximately -5 dB. When one ear was presented with $n(t) + s(t)$ and the other was presented with $n(t) - s(t)$, the SRT was approximately -8 dB, meaning that the speech was intelligible at lower relative levels. Thus, binaural headphone listening provided a "release from masking" of approximately 3 dB.

Another example of the power of binaural hearing is the spatial release from masking that occurs when the target and noise are located at different positions in space (Hirsh, 1950) or even different perceived positions in space (Freyman et al., 1999). A third example, noted qualitatively in 1931 by Von Békésy (1960, p. 376) and by Koenig (1950), and which is a kind of spatial release from masking, is the increase of the intelligibility of a single source in reverberation when listening with two ears instead of just one.

We now discuss some results in the psychoacoustics literature that are relevant to the intelligibility of speech when heard binaurally in reverberation and with interfering sources. It should be noted that the prediction of the intelligibility of anechoic speech mixed with stationary noise is very well described by the Articulation Index (AI) (French and Steinberg, 1947; Fletcher and Galt, 1950). This is a simple model based on the long-term spectral averages of the signal and interference. Some models of spatial hearing, hearing with speech interferers, and hearing in reverberation attempt to translate other types of noise into an effective frequency-dependent SNR that can be input to the AI calculation.

2.2.1 Single source in reverberation

Nábělek and Robinson (1982) reported the intelligibility of words presented in various reverberant conditions to subjects of various ages. Their results show that for subjects of all ages, reverberation decreased intelligibility, with greater amounts of reverberation (longer reverberation time) decreasing intelligibility more. In a six-way forced choice task, the subjects who performed the best, the 27 year old age group, when listening monaurally,

were correct 99.7% of the time for anechoic speech, but 97.0, 92.5, and 87.7% of the time for reverberation times of 0.4, 0.8, and 1.2 s, respectively. The 10-year-olds in this study performed significantly worse than the adults, suggesting that humans learn strategies for dealing with reverberation over the course of their childhood.

In the same study, Nábělek and Robinson (1982) also found that binaural listening improved intelligibility by approximately 5 percentage points as compared to monaural listening in all conditions. This provides a quantitative measure of the qualitative observations of the binaural release from reverberant masking observed by Von Békésy (1960) and Koenig (1950). Other measurements of binaural release from reverberation have found an improvement in intelligibility of 10 percentage points for phonetically balanced words (Moncur and Dirks, 1967) and 10-25 percentage points for the same six-way choice task (Nábělek and Pickett, 1974). While the values of these percentages vary with the task, it is clear that the intelligibility of reverberant speech increases under binaural listening.

Gelfand and Silman (1979) studied the effects of reverberation on the perception of specific phonetic information on rhyming words. They found that recognition performance decreased for both initial and final consonants, but decreased more for final consonants. In examining the effect of reverberation on types of phonetic information, they found that it interfered the most with the perception of place of articulation, and for stop and fricative consonants. The least affected phonetic information was sibilance, duration, and semivowel information. These results led the authors to conclude that reverberant interference acts much like additive noise, with greater noise coming after the more energetic vowels than after initial consonants. This explanation is known as overlap-masking.

The overlap-masking hypothesis has been qualified by Libbey and Rogers (2004), who measured the difference in masking caused by reverberation and by reverberation-like noise with the same time-frequency profile but randomized fine-structure. The randomized noise was combined with the direct-path signal in different direct-to-reverberant ratios and the intelligibility of phonetically balanced sentences was measured for diotic (same signal to both ears) and binaural presentations. For the real reverberation, binaural listening provided an advantage of 6 percentage points over diotic listening. For the reverberation-like noise, however, this advantage was only 2.6 percentage points. From this result, the authors concluded that the binaural release from masking can only partly be explained by a release from overlap-masking, while the rest must come from some other mechanism.

The reduction in intelligibility caused by reverberation of a particular channel can be predicted by the Speech Transmission Index (STI) (Houtgast and Steeneken, 1985). The STI is based on the so-called modulation spectrum of speech, which is essentially the Fourier transform of each filterbank band's envelope. Because reverberation smears out speech in time, it fills in troughs in the modulations of speech. By sending specific signals through a channel, the STI computes the effect of that channel on various modulation frequencies in various bands, from which it can predict the intelligibility of speech transmission in that channel.

2.2.2 Spatial release from masking

Spatial release from masking is the increase in intelligibility of speech that occurs when interference is moved from the same spatial location as the target to a different location. It can also be used to describe the difference in intelligibility of a target speaker in the presence of distracting noise or other speakers when the target and distractors are co-located versus being spatially separated. This effect is only significant when subjects are listening binaurally, and it occurs in both anechoic and reverberant environments.

Comprehensive studies on the spatial release from masking in anechoic environments include (Plomp and Mimpen, 1981; Bronkhorst and Plomp, 1988; Peissig and Kollmeier, 1997), which all examined the intelligibility of speech presented directly in front of the listener in anechoic conditions as a function of the azimuth of a noise masker. By definition the spatial release from masking is 0 dB at 0°, when the sources are colocated. They are in close agreement that the spatial release from masking is approximately 10 dB with the noise at 90° and 1–2 dB with the noise directly behind the listener. These results are shown in Bronkhorst (2000, fig. 2).

2.2.2.1 In reverberant environments

Prior to work in this decade, only a handful of studies have examined the effect of reverberation on the intelligibility of speech in a mixture (Hirsh, 1950; MacKeith and Coles, 1971; Plomp, 1976; Koehnke and Besing, 1996). We shall briefly discuss the findings of Plomp (1976), who measured the intelligibility of speech with a masker in both reverberant and anechoic conditions. In contrast to other studies, both speech and noise maskers were used in these experiments. All of the conditions placed the target utterance directly ahead of the listener at 0°, and in all but one condition the source-to-listener distance was 1 m. In the anechoic condition, Plomp found that the spatial release from masking speech was 4–5 dB, and as much as 6 dB with the masking utterance at 135°. As in the anechoic experiments described above, there was still a spatial release from masking when the masker was directly behind the listener of approximately 3 dB.

For the reverberant conditions, speech reception became more difficult for colocated sources as the reverberation time increased. When the source-to-listener distance was increased from 1 m to 2 m, the threshold increased further. The spatial release from masking (the difference between performance for colocated sources and spatially separated sources) was still present for all reverberation times, however, although it was smaller than in the anechoic case. For a moderate reverberation time of 0.4 s, spatial release from masking was approximately 3 dB at 135°, dropping to approximately 2 dB with the masker directly behind the listener. For longer reverberation times of 1.4 and 2.3 s, the spatial release from masking dropped to approximately 2 dB. And for the 2.3 s reverberation time and 2 m source-to-listener distance, it dropped to about 1 dB. Notably, for the longer reverberation times, spatial release from masking became approximately equal for all masker locations aside from 0°.

Plomp's results for the noise masker were very similar to those for the speech masker, but with an offset of 3 dB. This offset meant that the masking speech needed to be 3 dB louder than the masking noise to achieve the same level of intelligibility. This extreme similarity

is somewhat curious and has been revised by subsequent studies. It could be a result of the subjective methodology this study employed, in which the target-to-masker ratio was adjusted until listeners indicated that they believed the intelligibility of the sentence to be 50%.

Culling et al. (2003) performed related experiments, measuring intelligibility in a variety of conditions. The target and masker were either colocated at 0° or at ±60°, were either normally intonated, algorithmically modified to be monotonous, or algorithmically modified to invert the pitch contour, and were either simulated in an anechoic room or a reverberant room. They found that speech was more intelligible with natural intonation and with spatial separation in anechoic conditions, but not in reverberant conditions. This could be due to the relatively large source-to-listener distance (2 m) or relatively small room (40 m^3 in volume), both of which led to the low direct-to-reverberant energy ratio of −10 dB. They also created the head-related transfer functions used in the experiment with free-field microphones, creating only minimal interaural level differences, to eliminate any advantage for one ear cause by different ILDs (the "better ear advantage"). The lack of ILD might have interfered with listeners' abilities to use localization cues in separation.

2.2.2.2 Based on perceived location

An interesting result in spatial release from masking is that sometimes a true spatial separation is not necessary to release a target from masking, but only a *perceived* spatial separation. This is done using the precedence effect, the fact that humans generally localize a source based on its direct-path signal, described in more depth in section 2.3.2. Freyman et al. (1999, 2001) take advantage of this effect to manipulate the perceived position of the masking sound source without changing the amount of energy or the directions from which it comes. In the experiment, sounds are delivered to subjects in an anechoic chamber via two speakers, one directly ahead and one 60° to the right. For one type of stimuli, which we will refer to as"precedence effect" stimuli, an utterance was played from both speakers with a 4 ms delay between them. Because of the precedence effect, subjects perceived the sound as coming from the direction of the leading speaker. The target utterance was always presented so that it was perceived as coming from straight ahead.

Among other conditions, Freyman et al. (1999) examined the situation where both target and masker were these "precedence effect" stimuli with the target in center-leading presentation and the masker in either center- or right-leading presentation. When the masker was speech-shaped noise, there was little difference in terms of intelligibility between right-leading and center-leading masker presentation. When the masker was speech, however, there was a significant improvement in intelligibility for right-leading masker presentation compared to center-leading masker presentation. Even though the spatial distribution of speech energy was identical, the perception of a spatial separation was sufficient to improve intelligibility of the target.

Freyman et al. (2001) use the same experimental framework with a wider variety of masking signals including modulated speech-shaped noise, a mixture of two utterances, time-reversed speech, and speech in a foreign language the subjects did not understand. They showed that the release from masking due to perceived spatial location was still significant for non-noise stimuli, but that it was most pronounced in the situations where

the target and masker were most similar. The difference between the masking properties of speech and noise of the same root mean square (RMS) level will be discussed further when we discuss informational and energetic masking in section 2.2.3.3.

2.2.3 Streaming

One explanation for the results of these experiments on perceived spatial separation depends on the idea of *streaming* from auditory scene analysis (Bregman, 1990). Auditory scene analysis theorizes that the incoming soundscape is segmented using certain cues into small time-frequency fragments from the same "auditory object". After these fragments are formed, they are clustered with similar fragments using another, possibly different, set of cues.

In vision, space is a primary dimension, in that images on the retina are arranged according to the spatial position of objects in the world. In hearing, space is a computed dimension, arising from later processing of the signals from the two ears, while time and frequency are primary dimensions available at the very periphery of the perceptual system. The next sections discuss experiments that aimed to determine spatial hearing's influence on the initial formation of fragments and on the later streaming of fragments into auditory objects.

2.2.3.1 Source segmentation using spatial cues

A number of experiments have examined the role of space as a primary cue in auditory scene analysis. They have generally found that basic spatial cues like ITD and ILD only help segmentation weakly, but that perceived spatial location provides a strong cue for streaming (Darwin and Hukin, 2000a). Instead, spectro-temporal sound structure cues are thought to be the primary cues for segregation. These include common onsets and offsets, harmonic structure and continuity over time. Higher-order features are thought to be used to group segments together over time. These include timbre, pitch, vocal tract length, perceived location and even word identity, grammatical structure, and semantics (Shinn-Cunningham, 2008; Darwin, 1997).

According to Brown and Palomäki (2006), the idea that segmentation might occur before localization was first proposed by Woods and Colburn (1992) and pursued by Hill and Darwin (1996). Their hypothesis was that localization could be performed by pooling ITD over frequencies that other cues determined to be from the same source. These experiments involved presenting to subjects tone complexes with various combinations of cues, including binaural cues, and determining whether one or two objects were formed. They found that ITD alone was not sufficient to segregate concurrent sounds. These results have since been qualified by Darwin and Hukin (1999), who show that this isn't the case for more complex sounds.

Another experiment along these lines was reported by Culling and Summerfield (1995). In this experiment, four bandpass noise stimuli were used to simulate the first two formants of four vowels. Depending on their grouping, subjects would hear either "ah"+"ee" or "oo"+"uh". The authors showed that subjects were able to recognize these vowels when

a pair of formants was played exclusively to each ear over headphones. They were also recognizable when one pair of formants was coherent at the two ears and other was incoherent at the two ears. They were not recognizable, however, when pairs of formants shared their ITD in headphone presentation. These results imply that ITD does not allow the grouping of simultaneous signals.

The conclusion that this failure of ITD to facilitate segregation over headphones implies the failure of all spatial cues to facilitate segregation was qualified, however, by Drennan et al. (2003). They showed that when the same stimuli were presented in free-field listening conditions, the vowels were recognizable above chance levels. They also showed that they were recognizable from headphone presentation of these stimuli using full head-related transfer functions, which included ILD information. The authors hypothesize that the difference in performance might have come from the introduction of ILD cues, which would be more effective at cuing the location of the highest formant at 2 kHz, although this experiment did require much more training of the subjects (on the order of 30 hours each), and showed a wide inter-subject variability.

2.2.3.2 Source streaming using spatial cues

A number of papers have examined the role of higher-level cues in the streaming process (Darwin and Hukin, 1999, 2000a; Freyman et al., 2001; Brungart, 2001; Hawley et al., 2004). This process is tied to the idea of auditory attention (Ihlefeld and Shinn-Cunningham, 2008; Kidd et al., 2005a), which is thought to act on fragments already segmented from the incoming soundscape, and is thus involved in the streaming process. Ihlefeld and Shinn-Cunningham (2008) found that when attention is not focused on location, but on timbre, spatial separation still improves the segmentation of the incoming audio. When attention is focused on location, spatial separation improves both the segmentation and the selection of the proper segments.

Darwin and Hukin (1999, 2000a) studied the role of ITD, vocal tract length, and prosodic cues in streaming key words into cue sentences. In later work, Darwin and Hukin (2000b) performed the same experiment again, but in the presence of reverberation. In anechoic listening, pitch and other prosodic cues were not strong enough to override an interaural time difference of 91 μs, but when combined with a vocal tract length difference of $\pm$15%, they were able to override an ITD of 181 μs. In reverberation, however, listeners were less able to use ITD in streaming. Although they were able to use large ITDs to some extent, the vocal tract and prosodic cues were more robust to reverberation.

Brungart (2001) found that for diotic (same sound to both ears) mixtures of two speakers, intelligibility was highest for a speaker of the opposite gender, lower for a speaker of the same gender, and lowest for other utterances from the target speaker. Hawley et al. (2004) measured the binaural benefit of attending to speech with different types of maskers, and found that this benefit was larger for speech and time-reversed speech than for noise or speech-modulated noise. This result could be due to the release from informational masking in the case of speech, but not noise. Similarly, Freyman et al. (2001) found that the spatial release from masking was largest for speech in a language the subjects understood, smaller for speech in a language they didn't understand or time-reversed speech, and smallest for speech-shaped noise or speech-modulated noise, i.e. the interference power of

the masker tracks its similarity to the target speech. From a different perspective, at some level the maskers are all equivalent when they are spatially resolvable from the target, but as they become spatially inseparable from the target, their interference depends on their similarity to the target speech.

2.2.3.3 Informational masking

As mentioned previously, spatial release from masking is larger for a speech masker than for a noise masker. This is one phenomenon in a set of phenomena referred to collectively as *informational masking* (Durlach et al., 2003; Kidd et al., 2007). Masking in general is the effect that one signal has on another to decrease the fidelity of its perception. In the situations we consider here, this means that it interferes with the intelligibility of speech. The energetic masking of speech by stationary noise is a result of the noise blotting out the energy of the target signal, an effect which occurs at the peripheral auditory system. Informational masking, however, is defined as any masking that is not a result of energetic masking and by definition occurs higher up the processing chain. Because of this negative definition, there are likely a number of different effects currently grouped under this label.

There is one effect in particular that will be briefly discussed here, namely errors in streaming of segments of sounds. The type of errors subjects typically make when asked to attend to one speaker in a mixture of colocated speakers typically involve reporting information from the interfering speaker as opposed to random choices from those available (e.g. in a multiple-alternative forced-choice task). This has been shown for monaural listening with two speakers (Brungart, 2001) and for binaural listening (Ihlefeld and Shinn-Cunningham, 2008; Kidd et al., 2005a). This is less the case with non-native speakers of a language, however (Cooke et al., 2008). The implication of this result is that the short-term segregation process is functioning in these conditions, but the streaming together of these segments is failing.

Progress has been made towards studying both energetic and informational masking in isolation. Brungart et al. (2006) attempted to isolate energetic masking using the ideal binary mask, an idea originally from the automatic speech separation literature (see section 3.2 for further discussion). In this experiment, a mixture of 2–4 speakers was masked with the ideal binary mask for the target speaker and the intelligibility of the resulting separation was measured. The ideal binary mask was created by selecting all of the time-frequency points in the spectrogram for which the target source was a certain amount louder than the combination of the other sources. Their results showed that intelligibility was close to that of the target alone when the target-to-interference ratio was between -12 and 0 dB and dropped off outside of that region. Further experiments have shown, however, that this intelligibility gain includes more than just an elimination of energetic masking. In particular Wang et al. (2008) show that applying an ideal binary mask to *only noise* still yields quite intelligible results. Because the speech is intelligible with effectively infinite "energetic masking", additional information must be present in the shape of the mask itself.

Informational masking is more easily isolated. Using a cochlear implant front end, Arbogast et al. (2002) divided speech utterances into a number of frequency sub-bands. They compared the spatial release from purely informational masking with that from combined

informational and energetic masking when speech was masked by either another speaker or speech-shaped noise. In the purely informational masking condition, different utterances were assigned to disjoint sets of sub-bands. Because the utterances did not overlap in frequency at all, no energetic masking could occur. In the energetic-plus-informational condition, both signals were presented in the same bands. They found that for 50% intelligibility, spatial release from the informational masking was 18 dB for the speech masker, while it was less than 10 dB for the noise masker in both the informational and energetic-plus-informational masking conditions. This implies that spatial release from informational masking is larger for speech than for noise, and that informational masking and energetic masking for noise have approximately equivalent effects.

This experiment was extended to include the effects of reverberation by Kidd et al. (2005b). They found that when the masking was primarily energetic, reverberation decreased the spatial release from masking from 8 dB to 2 dB. When the masking was primarily informational, the spatial release from masking remained at 15–17 dB regardless of reverberation. This implies that while spatial release from energetic masking can be degraded by reverberation, spatial release from informational masking is more robust to reverberation.

2.2.4 Summary

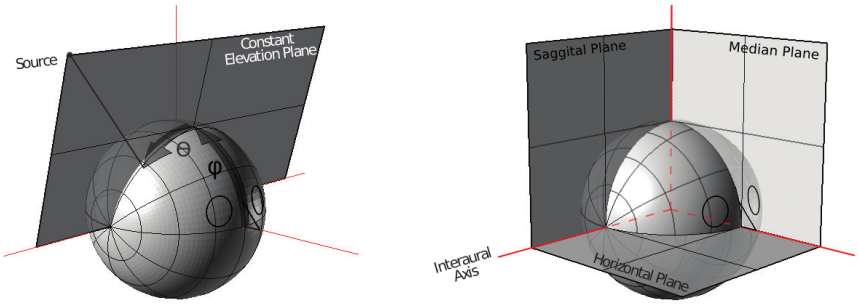
While its precise causes and consequences are still actively being investigated, it is clear that binaural hearing makes speech more intelligible in reverberation and in conditions where a target is spatially separated from maskers. Reverberation makes source separation more difficult, although it is more detrimental during monaural listening than binaural. Spatial cues appear to be involved less in the formation of segments of audio from one source than in the assembly of these segments into "auditory objects".

2.3 Human localization performance

Localization of sound by human listeners has been studied extensively for over 100 years. This section discusses only the points that are relevant to subsequent chapters including localization in anechoic environments, which has been very well studied; localization in adverse conditions, which has been well studied for certain conditions; and models of human sound localization. For a more complete treatment, the interested reader is directed to (Middlebrooks and Green, 1991; Grantham, 1995; Durlach, 2005; Stern et al., 2006; Blauert, 1997).

Before describing these results on localization, we briefly discuss the coordinate system used for discussing these results. This work uses the interaural coordinate system (Algazi et al., 2001b), shown in figure 2.3(a), which has symmetries similar to those of human auditory physiology. A point's position is specified by its elevation, azimuth, and range. In addition, some important features of this coordinate system are shown in figure 2.3(b): the interaural axis and the median, horizontal, and frontal planes.

Elevation is the angle between the horizontal plane and a point's projection onto the median plane. Azimuth is the angle between the a point's projection onto the plane at



(a) Interaural polar coordinates (b) Important features

Figure 2.3: The geometry of localization. A source is located at elevation ϕ and azimuth θ. Lines on the head show contours of constant azimuth and elevation.

that elevation that passes through the interaural axis and the median plane. Elevation measures the vertical angle of a point, ranging from 0° ahead, to 90° above, to 180° behind the listener. Below the listener is an elevation of both 270° and −90°. Azimuth measures the lateral angle of a point, ranging from 90° along the interaural axis to the right of the listener, to −90° along the interaural axis to the left of the listener. A surface of constant azimuth in this coordinate system defines a cone coaxial with the interaural axis.

Humans can localize sounds to varying degrees in azimuth, elevation, and distance. To a first order approximation, different cues are used for each dimension. Azimuth uses interaural time and level differences, as described by Lord Rayleigh a century ago (Strutt, 1907), elevation uses pinna cues, and distance uses a combination of direct-to-reverberation ratio, sound level, and high-frequency content (Grantham, 1995). The spatial resolution in azimuth or elevation of the auditory system is typically measured in terms of the minimum audible angle (Mills, 1958), which Blauert (1997) refers to as localization blur. This is the smallest angle separating two sources of the same sound that subjects can reliably discriminate.

2.3.1 Localization of a single source in an anechoic environment

Localization in azimuth (with which we are primarily concerned in this work) is based mainly on interaural time and level differences. Listeners tend to use ITD for frequencies up to 1–1.5 kHz, and ILD for frequencies above 4 kHz (Wightman and Kistler, 1992). Mills (1958) and Stevens and Newman (1936) showed that azimuthal localization performance is worse for tones between 2 and 4 kHz than at higher or lower frequencies, and that the ITD and ILD cues are both weak in that region.

In the best case, for a wideband source directly in front of the listener in anechoic conditions, azimuthal resolution can be as low as 1°. Makous and Middlebrooks (1990) found that averaged across listeners and conditions, this resolution was approximately 2°. Localization

resolution falls off for sources away from the median plane to approximately 10° for sources at 75° azimuth (Mills, 1958). Klumpp and Eady (1956) found that for wideband stimuli, the just-noticeable difference of ITD itself was about 10 μs. They found it to be similar for a 1 kHz tone, but much larger for tones much higher or lower.

Localization in elevation is mainly enabled by cues from the pinnae (outer ears), the folds of which create elevation-dependent filtering of incoming sounds (Roffler and Butler, 1968). It depends primarily on high frequencies, and although there is some debate over the lower cutoff frequency for these cues, Middlebrooks and Green (1991) estimate that frequencies above 4 kHz are important. Butler and Humanski (1992) showed that for sources off of the median plane, listeners also use interaural level and time differences to localize in elevation. Listeners learn to hear with their particular pinnae, and listening through headphones to a simulation using another person's pinnae leads to an increase in front-back and other confusions (Wenzel et al., 1993). Roffler and Butler (1968) found that filling in the folds of the pinnae prevented subjects from localizing sounds in elevation, although azimuthal localization was not affected. Hofman et al. (1998) found that listeners were able to relearn localization in elevation after several weeks of wearing molds that modified their pinnae, and were instantly able to localize sounds in elevation after the molds were removed.

(Grantham, 1995) reports that in the median plane localization in elevation is accurate to about 4°. Makous and Middlebrooks (1990) found that averaged across listeners and conditions, this localization blur was approximately 3.5°. This localization blur increases to 10° or more for sources slightly behind the vertical (Blauert, 1997, Fig. 2.5). Makous and Middlebrooks (1990) found that for some rear directions combined azimuthal and elevational localization blur was as high as 20°. Mills (1960) showed that the just noticeable difference in interaural level between two tones was between 0.5 and 1 dB.

Localization in distance is enabled by the direct-to-reverberant energy ratio, the source intensity, and the high frequency content of a signal (Grantham, 1995). The direct-to-reverberant ratio provides a cue to absolute distance in reverberant environments, as it decreases with increasing source-listener distance (Mershon and Bowers, 1979). Intensity aids distance perception because the energy in the direct-path signal decreases by 6 dB for each doubling of distance. Coleman (1968) found that subjects judged clicks with more high-frequency content to be closer to them, possibly because of the slight low-pass characteristic of air (Coleman, 1968), which could provide a cue for distance judgment over long distances. The familiarity of a source can transform a relative distance cue like intensity or high-frequency content into an absolute distance cue (Grantham, 1995; Mershon and Bowers, 1979).

Middlebrooks and Green (1991) discuss the possibility of localization with interaural phase and time differences for higher frequencies. While it does not appear that humans use these cues in such circumstances, it is possible that other animals might. Even though humans are only sensitive to interaural phase differences for sinusoids up to about 1.5 kHz, Rose et al. (1967) showed that there is significant phase coding in the auditory nerve of other primates up to 5 kHz. Moiseff and Konishi (1981) showed that the barn owl is sensitive to interaural phase differences for frequencies up to 7 kHz. At those frequencies, and because of the distance between the owl's ears, the IPD of a single tone is an ambiguous cue to localization, but IPDs at two frequencies provides enough information for the owl to resolve this ambiguity and infer an ITD (Knudsen, 1984; Middlebrooks and Green, 1991).

Another cue that is thought to be used in spatial perception is interaural coherence. Faller and Merimaa (2004) showed that cues from time-frequency regions of high interaural coherence provide more reliable localization than cues from other regions. Coherence can be used in this way to discriminate time-frequency regions dominated by direct-path signal as opposed to incoherent reverberation. Gabriel and Colburn (1981) showed that humans listeners are sensitive to changes in interaural coherence, much more so for high coherence signals than for low coherence signals. Coherence is measured on a scale from 0.0 to 1.0, and they found the just-noticeable difference for high-coherence signals to be 0.02, as compared to 0.3–0.4 for low coherence signals.

2.3.2 Localization in adverse conditions

Localization in reverberant environments has not been studied as thoroughly as localization in anechoic environments. In one early study, Hartmann (1983) measured subjects' ability to azimuthally localize sounds when sitting 12 m from a set of loudspeakers in a concert hall. He showed that reverberation did not affect the localization of tones with onset transients, which could be localized with an RMS error of about 3°, but that it made localization of tones with slow onsets nearly impossible. Broadband noise turned on slowly in the same way was localizable with a similar accuracy, although less so as the amount of reverberation increased.

Many more psychoacoustic experiments have examined the so-called precedence effect (Wallach et al., 1949), although these experiments typically only involve a direct-path sound and a small number of echoes, often just a single echo. The precedence effect involves a number of observations about localization in the presence of echoes, notably that early echoes are perceived as fused with the direct-path signal into a single auditory percept and the location of that percept is the location direct-path percept alone (Litovsky et al., 1999). The fusion of impulses into a single percept has been shown to build up over time, and will break down if certain characteristics of the leading and lagging stimuli are changed suddenly, and then build up again gradually (Clifton, 1987).

Localization accuracy has also been studied with multiple simultaneous sources. Good and Gilkey (1996) asked listeners to localize a square-wave stimulus in the presence of broadband noise. They found that localization accuracy decreased as SNR increased, but that some dimensions of localization were much more sensitive to the interference than others. In particular, at lower SNRs, listeners tended to make more front-back confusions, and localization in elevation was less accurate. Localization in azimuth was practically unaffected until the sound source was inaudible. Because it used an absolute localization task instead of measuring a minimum audible angle, and because of the coarse quantization of locations, Good and Gilkey (1996)'s results are quantitatively different from others discussed in this section. For localization in azimuth, RMS errors were approximately 12° in quiet, 18° at 0 dB signal-to-noise ratio (SNR), and almost 40° at −10 dB SNR. For localization in elevation, RMS errors were approximately 18° in quiet, 28° at 0 dB, and slightly less than 40° at −10 dB SNR.

More apropos to this work is the human ability to localize speech heard simultaneously with other speakers. Hawley et al. (1999) found that azimuthal localization is still accurate when listening to a mixture of three competing sentences and that intelligibility of speakers

in a mixture was not limited by localization ability. These localization results were similar to those of Good and Gilkey (1996) in that the localization task was an absolute task with coarse quantization. Also similar in task and in results was Yost et al. (1996), who found that listeners could localize up to three simultaneous speakers, even when listening to signals recorded at the ears of a KEMAR dummy head.

2.3.3 Models of human sound localization

Two main models of binaural localization have been extensively tested against the classic psychoacoustics literature, which was mainly concerned with headphone presentation of tones and noise. These are the cross-correlation model (Jeffress, 1948) and the equalization-cancellation (EC) model (Durlach, 1963). Many book chapters discuss these models in great depth and the interested reader is directed to (Stern et al., 2006; Stern and Trahiotis, 1995; Colburn and Kulkarni, 2005; Braasch, 2005).

The model of Jeffress (1948) and models derived from it localize sounds in azimuth using coincidence detection or cross-correlation. Jeffress proposed a neural mechanism for lateralization where individual neurons are coincidence detectors sensitive to particular interaural delays and the population of neurons can estimate the most likely delay. Sayers and Cherry (1957) were the first to attempt to relate this lateralization explicitly to the cross-correlation of the signals at the two ears, specifically using a running cross-correlation. Colburn (1973) instead based his model on the running cross-correlation of the signals as they would be represented on the auditory nerve, after having been processed by the peripheral auditory system. Most notably, this requires separate cross-correlations at different frequencies. Stern et al. (1988) further extended the model to include an explicit mechanism for converting cross-correlations into subjective judgments of laterality and added a coincidence mechanism to detect ITDs that were consistent across a number of frequencies.

The equalization-cancellation model was developed to explain the results of binaural detection experiments like those of Licklider (1948) performed over headphones with sinusoids in noise (Stern et al., 2006). The model proposes that the signal from one ear is equalized in amplitude and delay and then subtracted from the other signal to cancel the interfering source, hopefully revealing the target. This process is subject to internal noise and limitations on the accuracy of the equalization process, allowing it to render quantitative predictions. In the classic binaural detection experiment, when a sinusoid and white noise are both presented identically to the two ears, the sinusoid is detectable only at much higher levels than when it is inverted at one ear while the noise is still identical at both ears. The EC model predicts that in the first case the noise cannot be canceled without also canceling the sinusoid, whereas in the second case the noise can be canceled without affecting the sinusoid. This method can only be directly applied when a single, coherent noise source is present and fails in reverberation. It is similar to certain blind source separation systems in this way, as described in section 2.4.2.

A model that combines aspects of both the EC and cross-correlation models is that of Breebaart et al. (2001a,b,c). It uses a cancellation model but in a Jeffress-like coincidence framework. For each frequency band, the cancellation is performed at many interaural time and level differences simultaneously, creating a two-dimensional grid of responses.

The point or points at which the response is minimal, i.e. cancellation is maximal, provides an estimate of the interaural time and level differences.

2.3.4 Summary

Human localization performance is quite accurate, especially when the target sound is broadband or impulsive. It is relatively unaffected by reverberation and concurrent interfering sound sources. Although the classic models of human sound localization make accurate predictions for anechoic stimuli, they need to be developed further to make similar predictions in reverberation.

2.4 Separation systems

Many source separation algorithms have been proposed in the literature, although none of them can sufficiently separate a target speaker from a reverberant mixture well enough to restore automatic speech recognition performance to its levels in quiet. This section describes some of these approaches and their performance and limitations in reverberation. The broad categories of systems described include microphone arrays, blind source separation and independent component analysis, which typically utilize more than two microphones; monaural separators and strong prior models of speech and robust speech recognition, which typically utilize only a single microphone; and finally binaural computational auditory scene analysis (CASA) systems, which are most similar to this work, involving a pair of microphones. References to reviews of each of these topics are given in the appropriate sections, but for general overviews please see Weiss (2009, Chap. 2) and Brown and Palomäki (2006).

2.4.1 Microphone arrays

Initially developed for radar, sonar, and seismology applications, microphone arrays allow for accurate and powerful localization and separation of signals based on their direction of arrival. These initial applications were mainly concerned with narrowband signals, i.e. the bandwidth of the signal of interest was much smaller than its center frequency. When adapting microphone arrays to speech applications, however, these techniques had to be generalized to wideband signals, as even highly bandlimited telephone speech covers more than 4 octaves from 300 Hz to 3000 Hz.

The simplest narrowband beamformer is the delay-and-sum beamformer. When a microphone array is arranged in a linear structure, sound from a far away target source at a particular angle relative to the array arrives at each microphone in turn, delayed by an amount proportional to the distance between the microphones and dependent on the angle. By applying appropriate delays, all of the advanced signals can be brought into alignment in time, and summing them cancels out uncorrelated noise.

While the delay-and-sum beamformer can be applied to wideband signals, the width of the beam depends on the size of the array, measured in wavelengths, and so is frequency-dependent. This means that an off-axis wideband signal will not be rejected entirely, but only at certain frequencies, causing a net filtering effect. To remedy this problem, various array designs have been proposed, and Ward et al. (2001) describe a general design theory for constant beamwidth arrays. This theory treats space as a continuous quantity when constructing the array and only samples it with actual microphone positions at the very end. Additionally, instead of using a delay-and-sum technique, it uses the more general filter-and-sum technique, where for a given target direction each microphone is processed by a finite impulse response filter to impose an appropriate frequency-dependent delay.

These beamformers depend only on the position of a target source and are dependent on the calibration of the array while not being able to take advantage of the position of interfering sources. Adaptive beamforming methods address some of these shortcomings by focusing on both passing the target signal and rejecting unwanted interference and are less sensitive to array calibration. A narrowband adaptive beamforming technique was first introduced by (Capon, 1969) for use with seismic data. This method aims to pass signals coming from the "look" direction unmodified while minimizing the total output power, which has the effect of canceling signals from all non-look directions. While this method was successful, it involved a matrix inversion that was computationally expensive and numerically sensitive.

A wideband generalization of Capon's adaptive beamformer was introduced by Frost (1972), which applies an adaptive filter to each microphone signal, allowing frequency-dependent directional cancellation. This method was slow to converge and an improvement on it was presented by Griffiths and Jim (1982), which transformed the linearly constrained minimum variance problem into a constraint-free minimization, known as the generalized sidelobe canceler. This method is still slow to converge, but is simpler than Frost's method and has become the most common adaptive beamformer (Feng and Jones, 2006).

Because of their linearity, all of these algorithms can be thought of as steering nulls in space. Some of them can direct these nulls independently at different frequencies. This property is due to the fact that at each frequency, microphones are combined with gains and delays, making them equivalent to a finite impulse response filter, i.e. an all-zero filter.

2.4.2 Blind source separation

Adaptive beamformers can be considered one class of solutions to the general problem of blind source separation (BSS) (Douglas, 2001; Pedersen et al., 2007). The term "blind" serves to emphasize that neither the signals nor the mixing system (array parameters) are known by the separator. While beamforming focuses on the spatial qualities of the sources, independent component analysis (ICA) is another class of BSS algorithms that take advantage of the statistical properties of the signals themselves.

ICA was first proposed by Bell and Sejnowski (1995), who focused on solutions to the instantaneous mixing problem, i.e. the measured signals were assumed to be noiseless linear combinations of a set of input signals. By optimizing appropriate signal-dependent criteria, the original signals could be unmixed from one another up to an arbitrary scaling

and permutation of the sources. Many overviews of ICA have been written, the interested reader is referred to (Hyvärinen, 1999; Choi et al., 2005).

Because acoustic mixtures tend not to be instantaneous, but include delays and convolutions with spatial impulse responses, Smaragdis (1998) proposed applying ICA to individual frequency bands of spectrograms instead of directly in the time domain. Thus any delays or convolutions that were shorter than the window used to calculate the spectrogram were transformed into phase modifications and instantaneous ICA could still be applied. Solving a separate ICA problem in each frequency leads to the so-called source permutation problem, in which the wideband signals must be reconstructed by combining the appropriate narrowband separations. Many methods have been proposed to resolve this problem including comparing the envelopes of the separated narrowband signals (Ikeda and Murata, 1999), using the spatial location of the narrowband signals (Saruwatari et al., 2001), and a combination of the two (Sawada et al., 2004).

BSS has also been generalized to longer convolutions using a number of approaches. Pedersen et al. (2007) provide an excellend overview. Convolutive BSS systems take advantage of both spatial and signal-based properties of sound mixtures. One particular example of a convolutive ICA system is Triple-N ICA for Convolutive mixtures (TRINICON) (Buchner et al., 2004). It takes advantage of the non-whiteness, non-stationarity, and non-Gaussianity of speech to estimate a convolutive unmixing system. Extensions of this work have been published in (e.g. Buchner et al., 2005; Aichner et al., 2007) and will be discussed in chapter 5.

All of the microphone array and BSS techniques discussed so far can only separate sources from so-called *overdetermined* mixture conditions. This means that there are at least as many microphones as there are sources to be separated. They cannot separate sources from *underdetermined* mixtures where there are more sources than microphones. In this work, we present methods designed for separating sources from underdetermined mixtures, as humans are able to do.

2.4.3 Speech models and monaural speech separation

When some information about the target signal is known, it is possible to improve separation performance. Such information is especially important in the most underdetermined mixture possible, a monaural mixture, i.e. a single-microphone recording of multiple sound sources.

In the speech recognition literature, speech is modeled using a hidden Markov model (HMM) in which each state represents a small piece of a phoneme and hundreds of thousands of these small pieces are strung together in a very sparse network. The acoustical model at each state captures much of the variability in the instantiation of any particular observation while the network captures constraints at the acoustic, phonemic, word, and language levels. The transcript generated by a speech recognizer depends on the path that a particular utterance traverses through this network of connections.

Because speech enhancement and source separation systems do not need to generate a transcript, they have more freedom in the models that they use. Many ignore temporal

constraints altogether and focus simply on modeling the acoustic observations with Gaussian mixture models (Kristjansson et al., 2004; Reddy and Raj, 2004; Roweis, 2003). Others, however take advantage of at least simple temporal constraints (Ephraim et al., 1989; Varga and Moore, 1990; Gales and Young, 1996; Weiss and Ellis, 2007; Reyes-Gomez et al., 2004).

Early approaches to model-based source separation focused on speech recognition in the presence of noise (Varga and Moore, 1990; Gales and Young, 1996). Varga and Moore (1990) proposed the use of factorial hidden Markov models, in which both the speech and the noise are modeled with an HMM and observations are compared to the joint state space of both models. They also proposed the use of the so-called log-max approximation, which approximates the sum of two signals' power (measured in dB) as the maximum of the two. Gales and Young (1996) use a similar approach known as parallel model combination with explicit speech and noise models. This technique allows their system to increase the vocabulary of their system from 10s of words to 1000 words. Another early system, Ephraim et al. (1989) focused on speech enhancement instead of recognition by alternately estimating parameters for speech and noise models.

More recently, authors have been able to relax temporal constraints by using more complicated models of the interaction of the mixed signals (Kristjansson et al., 2004) or by using clever pruning techniques to make larger models tractable (Roweis, 2003). HMMs with larger state spaces can also be used if inference is carried out in separate frequency bands and coupled using a loopy belief propagation algorithm (Reyes-Gomez et al., 2004). And more complex speaker models can be used if parameters only need to be learned in a lower-dimensional parametrization (Weiss and Ellis, 2007). Larger speech models can also be combined with microphone array processing to improve speech separation (Attias, 2003).

Finally, Hershey et al. (2009) show that a system using the sophisticated observation combination technique of (Kristjansson et al., 2004) along with temporal constraints both on the short-term acoustic scale and the long-term word and grammar scale, can outperform humans in certain circumstances. In particular, they outperformed human subjects in the monaural speech separation challenge (Cooke et al., 2009), which used a very constrained grammar. By limiting their recognizer to that grammar, Hershey et al. (2009) were able to achieve an error rate of 21.6% compared with the human error rate of 22.3%.

2.4.4 Missing data speech recognition

While many of the previous systems integrated the speech separation process into the recognition process, another research track has explored the recognition of speech from incomplete observations of an utterance (Raj et al., 2000; Cooke et al., 2001; Raj et al., 2004; Palomäki et al., 2004; Kolossa et al., 2005; Barker et al., 2005). These systems, known as missing data speech recognizers, are typically used in tandem with other source separation algorithms, which provide an estimate of the specific times and frequencies that belong to the target source.

Three approaches to the process of missing-data speech recognition have been proposed based on the idea that a time-frequency mask of reliability is available for the target speech. The first reconstructs the speech in the regions of the spectrogram that are determined

to be unreliable (Raj et al., 2000, 2004). The second computes the mel frequency cepstral coefficients of the corrupted speech while converting uncertainty in the spectral domain into the cepstral domain (Kolossa et al., 2005). The final method performs recognition only on the reliable regions of the spectrogram while taking into considering constraints on unobserved parts of the target signal created by observed parts of the interfering noise (Cooke et al., 2001; Palomäki et al., 2004; Barker et al., 2005).

2.4.5 Binaural CASA systems

Many systems have been proposed to take advantage of the cues that humans use to perform "auditory scene analysis" (Bregman, 1990) in a process referred to as computational auditory scene analysis (CASA). While many CASA systems were monaural, a few have been binaural. These typically combine ideas of source localization with time-frequency masking, as our system does.

Perhaps the earliest was presented by Lyon (1983), who proposed using the cross-correlation of the outputs of binaural cochlear filterbanks to localize and separate sounds. The system was very computationally intensive at the time, and he was only able to run it on a single 200 ms example. It was able to separate a vowel sound from an impulsive sound and their attendant reverberations, although because the microphones were very close together there was little problem with the multimodality of the narrowband cross-correlations.

A similar system was proposed by Bodden (1993) that additionally included a model of interaural level difference. After estimating the azimuth of each source, it derived a soft mask for separating the target source from interference. According to Brown and Palomäki (2006) this system did not perform well in reverberation because strong early reflections were classified as separate sources. An extension of this work that uses interaural coherence to distinguish between direct-path and reverberant sounds is described by Kollmeier et al. (1993) and Wittkop et al. (1997). Aoki and Furuya (2002) have extended these systems further to explicitly model of the dependence of IPD and ILD on direction and weighting binaural cues by their reliability. (Liu et al., 2001; Palomäki et al., 2004) are in this vein as well.

While these systems may perform relatively well and include many ideas related to MESSL, their construction is rather heuristic. chapter 3 examines the problem from a more statistically-minded point of view, on which MESSL is based.

2.4.6 Summary

Many systems have been presented to approach the problem of understanding or isolating speech in noise. These systems run the gamut from very plausible binaural systems to exercises in linear algebra. Many place limitations on the number of sources that they can separate based on the number of microphones available, although this is not a problem for time-frequency masking techniques. Many are tuned for a particular type of noise or signal and fail for other signals, although this is not a problem for systems that model the interaural parameters instead of the signals directly. Many are not robust to

reverberation or coherent reflections, although this is not a problem for systems that place looser restrictions on the relationships between frequencies.

2.5 Localization systems

While many localization systems have been developed in the radar, sonar, and seismology communities, they typically rely on large arrays of sensors. Because in this work we are concerned with the localization and separation of sources using only two sensors, we will not devote much space to these array techniques. A two-sensor array limits the performance of classical localization techniques as do reverberation and interfering sources. Two-microphone methods almost exclusively focus on localization in azimuth, ignoring elevation and distance.

When only two microphones are available, cross-correlation is a popular localization technique. Systems based on cross-correlation are similar to the models of binaural localization described in section 2.3.3. In particular, cross-correlation has been studied extensively as a method for machine localization of sounds as well. Knapp and Carter (1976) describe the generalized cross-correlation (GCC) framework in which a number of localization algorithms fit. These methods produce point estimates of the time delay between two microphones by including a weighting function in a cross-correlation calculation. Under the assumption of uncorrelated, stationary, Gaussian sources, the maximum likelihood estimate of the delay fits into this formulation. Another particularly useful localization algorithm in this framework is the Phase Transform (PHAT), which whitens the two signals before cross-correlating them to provide a more sharply peaked correlation. While these methods generally perform well in uncorrelated noise, reverberation leads to increased bias, variance, and spurious detections (Champagne et al., 1996).

Aarabi (2002) extended this method to the localization of multiple speakers by calculating point estimates of source localization using the generalized cross-correlations on small time scales. These estimates are then pooled together over longer time scales in a histogram and the peaks of that histogram can be used in localizing multiple simultaneous sources. The rationale behind this two-stage approach is that each source will dominate the GCC performed in some of the frames at the shorter time scale. While this is typically the case for speech, it may work less well when sources overlap more in time, but less in frequency as in certain musical signals. Empirically, this technique also works well in reverberation.

A method that is very similar to our own is presented by Nix and Hohmann (2006) and based on histograms of interaural phase and level differences. Their system was able to successfully localize sounds in azimuth and elevation in the presence of realistic noise conditions, although the target speech was anechoic. While our own method is parametric, Nix and Hohmann (2006) use nonparametric estimates of source parameters, which can be more accurate, but require much more data to build.

One more binaural localization system worth mentioning is that of Wilson and Darrell (2006), which learns in a data-driven way to predict the reliability of localization cues in time-frequency regions of a signal from the magnitude spectrum. By passing the same signal through a large number of simulated room impulse responses, the reliability of

localization cues can be estimated. A regression of this reliability on the magnitude spectrum learns filters that can predict this reliability from novel speech. The predictors that the system learns show features of the precedence effect, in that both broadband power and an onset localized in frequency lead to higher predictions of reliability. While this reliability predictor was used for localization, it is related to those that are used for separation in missing data speech recognition discussed in section 2.4.4.

Microphone array techniques Dibiase et al. (2001) categorizes microphone array localization techniques as belonging to either cross-correlation or time-delay methods, steered response power (SRP) systems, or high-resolution spectral estimators. Cross-correlation methods have already been discussed, but it should be noted that with N microphones, cross-correlations can be performed between all $N(N-1)/2$ microphone pairs and then a physical location can be determined intersecting the set of hyperbolas generated by these time differences.

Localization methods based on the steered response power are more computationally intensive, but tend to be more robust to reverberation and to be able to extract more information out of the same number of observations (Dibiase et al., 2001). In these methods, a microphone array is steered to search for the point in space with the most power. These methods typically assume that the target and interference sources' spectra are known and that there is no reverberation, although Dibiase et al. (2001) describe their own SRP-PHAT algorithm, which is robust to reverberation. In the two-microphone case, SRP reduces to a simple cross-correlation.

The final class of microphone array localization algorithms are the high-resolution spectral estimators, including MUSIC (Schmidt, 1986) and ESPRIT (Roy and Kailath, 1989). These methods analyze the eigen-structure of the matrix of spatial correlations between microphones and with N microphones can localize $N-1$ sources. Correlated noise from early reflections tends to distort the correlation matrix, leading inaccurate localization. These algorithms were originally designed for narrow-band signals, but have been extended, at much greater computational expense, to wideband signals.

2.6 Summary

This chapter has provided some background on the tasks of sound source separation and localization. Humans perform well at these tasks, especially in interfering noise and reverberation. Automatic systems have just started to reach comparable performance on very limited versions of these tasks.