Chapter 3

Statistics of interaural parameters

This chapter presents an overview of the behavior of the signals in the source separation problem. Section 3.1 builds on an extended example of two signals mixed in reverberation and examines the interaural parameters of each of the sources in isolation and in the mixture. It explains the beginnings of the model of interaural parameters that will be used in subsequent chapters and it examines the effect that reverberation and early echoes have on the interaural parameters of individual sound sources. Section 3.2 describes a number of source separation masks that can be derived from pre-mixed sources, each of which can be considered optimal in some sense. These masks will also be used in evaluations in subsequent chapters.

3.1 Interaural cues

As discussed in section 2.3, humans use a number of cues for localizing sound sources, namely the interaural time and phase differences, the interaural level difference, the direct-to-reverberant ratio, spectral cues from the outer ears (pinnae), and the coherence of the signals at the two ears. This work focuses on the interaural time, phase, and level differences and occasionally mentions the coherence and direct-to-reverberant ratio because it is primarily concerned with localizing sources in azimuth. While section 2.3 discussed localization in each of the three dimensions of polar coordinates, this section in particular focuses individually on the three interaural cues used for localization in azimuth, interaural time, phase, and level differences. See figure 3.1 for a schematic representation of the causes for these cues, which will now be described in greater depth.

The interaural time difference (ITD) is caused by the finite velocity of sound. When a sound source is located closer to one ear than another, there is a delay between the signal’s arrival at the two ears. For a human listener, ITDs are generally bounded by $\pm 750 \mu s$ (Algazi et al., 2001b), although this limit depends on the size and shape of the listener’s head. The average human head width is approximately 145 mm (Algazi et al., 2001b), although with a speed of sound of 340 m/s, an interaural delay of 750 $\mu s$ corresponds
Figure 3.1: The causes of interaural time, phase, and level differences. From Ellis and Mandel (2009)

to an acoustic distance of 255 mm. This extra distance is due to the sound following the circumference of the head instead of the diameter at higher frequencies.

Because of the symmetry of the head and ears, the ITD conveys information about the azimuthal location of a sound source. For a spherical head, a fixed ITD between two ears defines a hyperboloid of possible source locations, which can be approximated as a cone far from the head. This cone is known as the “cone of confusion” (Mills, 1958) because ITD alone cannot resolve the location of a sound source on such a cone. In the interaural coordinate system described in section 2.3, such cones are very easy to describe, as they have a constant azimuthal angle. Humans resolve this ambiguity using pinna cues. The complicated folds of the outer ears (pinnae) impart a coloration to a sound that depends on the direction of the source relative to the listener’s head in both azimuth and elevation.

Interaural phase difference (IPD) is closely related to ITD, but more appropriate for narrowband signals. Narrowband signals are similar to sinusoids. Because such signals are close to being periodic, it is very difficult to distinguish a delay of more than one cycle from the corresponding delay of less than one cycle. This ambiguity can be considered a form of aliasing, and it is known as spatial aliasing when it is due to the spatial configuration of microphone arrays. It is particularly relevant when working with analyses of filterbank channels or other spectral representations. In such analyses, it makes more sense to analyze the interaural phase difference (IPD) as opposed to the interaural time difference.

The interaural level difference (ILD) is caused by the “shadowing” of the far ear by the head. Because a listener’s head is large relative to certain wavelengths of sound, it serves as a barrier to them, creating a shadow. If the diameter of the head is 145 mm, then it is approximately the size of one wavelength for sounds at 2300 Hz. Thus the head is much larger than the wavelength of sounds above 3–4 kHz and it serves as an effective obstruction to them. For sounds with wavelengths comparable to the head, the head acts as a diffuser, scattering these sounds when they reach the head. For sounds with wavelengths much larger than the head, the head is effectively transparent, and they pass around it undisturbed. The net effect of this is that the shadow and ILD only exists for sounds above
3.1.1 Example mixture

This section presents an example reverberant mixture of two people speaking and describes some of the properties of the mixture and of the features extracted from it that are used in subsequent chapters. The target speaker is female and is located at 0°, saying, “Presently, his water brother said breathlessly.” The interfering speaker is male and is located at 75°, saying, “Tim takes Sheila to see movies twice a week.” These utterances come from the TIMIT acoustic-phonetic continuous speech corpus (Garofolo et al., 1993), and the impulse responses that spatialize them were recorded in a real classroom and come from (Shinn-Cunningham et al., 2005). This same example mixture is used in section 5.3 to show example masks and parameter estimates of both our system and the other systems evaluated in chapters 5 and 6. Sound files from this example are also available on the project’s webpage.

\[\text{http://labrosa.ee.columbia.edu/projects/messl}\]
3.1 Interaural cues

Figure 3.3: Example observations of the interaural spectrogram: IPD and ILD. Two speakers in a reverberant classroom, the target at 0° and the masker at 75°.

Figure 3.2 shows the basic observations on which all of the systems discussed in this work are run. It includes the left and right channels of the individual sources before mixing and of the mixture. These plots only show the magnitude (in dB), but do not include phase information. Because the timing differences between the two ears are much smaller than the window size used in making the spectrograms, the timing information visible in the spectrogram (i.e. the horizontal alignment of the images) appears to be identical. The relative level of the two ears is more noticeably different, both in the individual signals and in the mixture. While the target has 0 ILD because it is directly in front of the listener, the masker is closer to the listener’s right ear, so more energy is apparent in its plot in the bottom row than in the top row.

The mixture also shows the extent to which the details of each source survive the mixing, at least coarsely. Even though many faint harmonics are noticeable in the target’s spectrograms, they are generally not present in the mixture, having been overwhelmed by energy from the masker. Some of the target’s higher power harmonics, however, are still noticeable at the low frequencies.

Figure 3.3 shows the observations as the interaural spectrogram, which is displayed as its magnitude and phase components, the interaural phase and level differences, respectively. The IPD is shown in the top row of figure 3.3. The position of the target again leads...
to a relatively uniform IPD across frequency, especially in regions of high energy. The IPD regions of low energy is quite evenly distributed throughout $[-\pi, \pi]$. The masker’s IPD, on the other hand, changes quite rapidly with frequency. It completes a $2\pi$ cycle approximately every 1.7 kHz, the first three of which are very clearly visible as a “rainbow” pattern. At higher frequencies, where the masker is less energetic, the pattern becomes more difficult to discern. In the mixture, both of these IPDs are well preserved. The blue regions from the target are distinguishable from the rainbow regions of the masker.

The ILD is shown in the bottom row of figure 3.3. The ILD of the target is close to 0 throughout because target is positioned directly in front of the listener. In time-frequency regions where there isn’t much direct-path energy from the target, however, the variance of the ILD increases. The outlines of these high energy regions are most visible at frequencies between 2 and 6 kHz, where the target is most energetic. Reflecting the level difference observable in figure 3.2, the masker’s ILD is quite large at high frequencies. Because of the frequency dependence of head shadowing, described in section 3.1, this ILD is strongest above 4 kHz, although it is noticeable down to 2 kHz. The ILD of the mixture nicely reflects the ILDs of the two individual sources, with only a small amount of interaction between them.

3.1.2  Model of interaural cues

This section begins the discussion of a model of the interaural parameters of a single source and its relationship to the above observations. It is further developed for localization in chapter 4 and for separation in chapter 5. Denote the sound source as $s(t)$ and the signals received at the left and right ears as $\ell(t)$ and $r(t)$, respectively. For a sufficiently narrowband source, the two received signals relate to the source by some delay and gain, in addition to a disruption due to noise. For a wideband source, this delay and gain can vary with frequency, which, in the time domain, manifests as a short impulse response at each ear. These impulse responses capture information about early echoes. In subsequent chapters we explore both frequency-dependent and frequency-independent models of this delay and gain.

For analytical convenience, we assume a noise process that is convolutive in the time domain, making it additive in both the log-magnitude and phase domains. For the frequency-independent model, the transfer function is modeled as a single large, deterministic coefficient at a certain delay and small, randomly changing coefficients at all other delays. The frequency-dependent model does not have as simple a time-domain interpretation. As discussed in chapter 4, a reverberant noise model is still able to localize sources in the presence of additive noise, the noise model typically used by other authors.

Combining the frequency-dependent gains and delays into two short impulse responses, $g_\ell(t)$ and $g_r(t)$, the various signals are related by:

$$\ell(t) = s(t - \tau_\ell) * g_\ell(t) * n_\ell(t) \quad r(t) = s(t - \tau_r) * g_r(t) * n_r(t). \quad (3.1)$$
The ratio of the short-time Fourier transforms, \( \mathcal{F}\{\cdot\} \), of both equations is the interaural spectrogram,

\[
\frac{L(\omega,t)}{R(\omega,t)} = 10^{\frac{\alpha(\omega,t)/20}{e^{j\phi(\omega,t)}}}
\]

\[
\approx 10^{\frac{\alpha(\omega)/20}{e^{-j\omega \tau(\omega)}}} N(\omega,t)
\]

where \( N(\omega,t) = \frac{N_r(\omega,t)}{N_i(\omega,t)} = \frac{\mathcal{F}\{n_r(t)\}}{\mathcal{F}\{n_i(t)\}}, \) \( \tau(\omega) = \tau_r - \tau_t + \omega^{-1} N(\omega) \), \( a(\omega) = 20 \log_{10} |G(\omega)| \), and \( G(\omega) = \frac{\mathcal{F}\{g(t)\}}{\mathcal{F}\{g_r(t)\}} \). Equation (3.2) is the ratio of the actual observations at both ears, while equation (3.3) is our model of that ratio. The two are not equal because the model assumes that any energy that cannot be explained by the frequency-dependent gain and delay is part of the noise.

Equation (3.2) factors the interaural spectrogram into \( \phi(\omega,t) \), the interaural phase difference (IPD) at frequency \( \omega \) and time \( t \), and \( a(\omega,t) \), the interaural level difference (ILD) measured in dB. Equation (3.3) models a source at a particular location with the frequency-dependent interaural time difference (ITD), \( \tau(\omega) \), and the frequency-dependent interaural level difference, \( a(\omega) \). Note that when the interaural time difference in frequency independent, it is just a single delay, which becomes a linear phase advance in the frequency domain

\[
\mathcal{F}\{x(t - \tau)\} = e^{-j\omega \tau} \mathcal{F}\{x(t)\}.
\]

Because the IPD is constrained to be in \([ -\pi, \pi ]\), even the frequency-dependent model never gets very far from this frequency-independent trend.

As will be demonstrated in the next section, \( a(\omega) \) and \( \tau(\omega) \) capture the spatial information from the direct-path and early echoes, while \( N(\omega,t) \) captures the spatial information from the late reverberation along with the possibly inconsistent information from previous time frames.

For this model to hold, \( \tau \) must be much smaller than the window over which the Fourier transform is taken. For dummy head or in-ear recordings of people, position-dependent delay differences are limited to \( 750 \) ms, while the window length used in these experiments is almost 100 times bigger (1024 samples at a sampling rate of 16 kHz). Similarly, \( g(t) \) must be smaller than the window, but because distinguishing between \( g(t) \) and \( n(t) \) is an ill-posed problem, parts of \( g(t) \) beyond one window’s length can be considered part of \( n(t) \), with a corresponding increase in the noise variance.

Localizing a signal requires inference of \( a(\omega) \) and \( \tau(\omega) \) from the observed interaural parameters \( a(\omega,t) \) and \( \phi(\omega,t) \), while minimizing the effects of the noise, \( N(\omega,t) \). In order to accomplish this goal, however, it is necessary to characterize the behavior of the noise in both magnitude and phase for known \( a(\omega) \) and \( \tau(\omega) \). Because the noise is modeled as convolutive, it is possible to analyze its effects on phase and magnitude separately

\[
\log |N(\omega,t)| = a(\omega,t) - a(\omega) \quad \angle N(\omega,t) = \phi(\omega,t) + \omega \tau(\omega) + 2k\pi,
\]

where the integer \( k \) is chosen so that \( \angle N(\omega,t) \in [ -\pi, \pi ] \), due to the inherent \( 2\pi \) ambiguity in phases. We will call these observations with the main trend subtracted the ILD and IPD residuals.
3. Statistics of interaural parameters

The next section explores the effects of the direct-path, early echoes, and late reverberation on the $a(\omega)$, $\tau(\omega)$, and $N(\omega, t)$ terms. It shows that the direct-path and early echoes mainly affect the $a(\omega)$ and $\tau(\omega)$ terms and that the later reverberation mainly affects $N(\omega, t)$. It also shows that it is reasonable to assume that the noise is independent and identically distributed across time with a relatively constant variance across frequency.

3.1.3 Interaural noise observations

This section explores the effects of various parts of the binaural room impulse responses on speech in order to validate the model described in the previous section. It does this by examining two-dimensional histograms of IPD or ILD as a function of frequency, which summarizes these observations over time. To generate one of these plots (e.g. figure 3.4), an utterance is convolved with a specific part of a binaural impulse response and then converted to the IPD and ILD representation of equation (3.2). This section also compares the interaural parameters of these utterances to those of utterances generated by adding speech-shaped Gaussian noise to anechoic speech signals. The effects of additive noise are similar to those of reverberation and fit in this modeling framework, although there are notable differences.
Figure 3.4 shows two-dimensional histograms of the IPD and ILD at different frequencies for two different sources, revealing the basic dependence of IPD and ILD on azimuth. The plots on the left side of this figure show that both the IPD and ILD of the target source at 0° are close to 0. There are certain frequency-dependencies in the means and variances of the lines, which are most noticeable at high and low frequencies, where the energy of this source is lowest.

The plots on the right side of the figure show that the IPD and ILD for the masker at 75° deviate substantially from 0. The IPD shows a basic linear trend, wrapping around 2π five times, indicating a delay of approximately −10 samples or −625 µs. The variance of the IPD is relatively constant, and comparable to the variance for the source at 0°, except at high frequencies, where it is much larger. The ILD is very negative, indicating that the signal is 20 dB louder on the right than on the left at frequencies above 5 kHz, and approaches that ILD approximately linearly from lower frequencies. The ILD variance is similar for the two sources, but slightly larger for the source at 75°.

### 3.1.3.1 IPD noise

The IPD and IPD residual are shown in figure 3.5 for a signal convolved with the same BRIR truncated at different points. In the first plot the source is convolved with only the direct path. In the second plot it is convolved with the sum of the direct path and early echoes. And in the third plot it is convolved with the whole impulse response, including the late reverberation. The top row shows a histogram of the IPD of the resulting spatialized signals. At each frequency, these histograms show the IPD to be unimodal, with roughly the same mean for all of the impulse response lengths, but with increasing variances. When the impulse response is just the direct path, the high frequencies are very well resolved and appear to fit the basic trend of the low frequencies. With the longer impulse responses, however, the high frequencies get “washed out” to some extent, and the trend is less easily discernible. The IPD mean at the high frequencies also changes rapidly for these signals, probably due to the influence of the early echoes.

The bottom row of plots in figure 3.5 shows the IPD residual, i.e. \( \phi(\omega,t) \) with \( \omega \tau(\omega) \) subtracted out of it. In these plots, \( \omega \tau(\omega) \) is the mean of the direct-path IPD at each frequency, leading to a perfectly flat line in the direct-path residual plot. The other two plots are also centered at 0, meaning that they follow the same trend line, but the variance of the residual is much larger. Especially for the signal convolved with the full impulse response, there are points at all possible values of the IPD residual, although they are concentrated at the mean. Again, the most reliable regions are the frequencies from 0 to 5 kHz, beyond which the variance increases substantially.

Note that the IPD of the target and masker in figure 3.3 are the same at harmonically related frequencies. This is most notable in (c), the IPD mixture plot, where it occurs around 1.6 kHz and 3.2 kHz. This happens because of the 2π wrapping of the phase of the masker, as can be seen in figure 3.4.

As shown in equation (3.4), the dominant relative delay between the two ears manifests itself in the frequency domain as a linear phase advance with frequency. For any two delays for the target and masker, \( \tau_1 \) and \( \tau_2 \), the resulting IPDs will then be \( e^{j\omega \tau_1} \) and \( e^{j\omega \tau_2} \).
3. Statistics of interaural parameters

![Graph of IPD and IPD residual](image)

**Figure 3.5:** Two-dimensional histograms of IPD and IPD residual of a source at 75°. The IPD residual is computed by subtracting the mean of the direct-path IPD out of the IPDs. The same impulse response is used for the three conditions, but it is truncated at 10, 32, and 565 ms.

These are equal for all frequencies at which

\[
\omega \tau_1 \equiv \omega \tau_2 \mod 2\pi \quad (3.6)
\]

\[
\omega = \frac{k2\pi}{\tau_1 - \tau_2} \quad (3.7)
\]

for all integers, \(k\), i.e. every \(\frac{2\pi}{\tau_1 - \tau_2}\). This effect can also be seen on the right side of figure 4.8 where the two lines cross.

**3.1.3.2 ILD noise**

The ILD noise is shown in figure 3.6 for the same utterance convolved with the various parts of the same impulse response. The top row shows the ILD corresponding to the signals in the top row of figure 3.5, while the bottom row of figure 3.6 shows the ILD of just the early echoes and just the late reverberation.
As can be seen in figure 3.6(d), the ILD trend in the early echoes is opposite that of the direct-path, i.e. the ILD is positive for the early echoes, but negative for the direct-path. This makes sense if the early echoes’ ILD is dominated by a virtual source on the opposite side of the body from the true source. When the two signals are added together, the two trends are also added, as can be seen in (b).

The basic trend in the ILD plots is that as more of the impulse response is used, the ILD shrinks towards 0 dB. At 7 kHz, the direct-path ILD is close to $-30$ dB. When early echoes and late reverberation are introduced, however, the mean ILD at 7 kHz decreases to $-15$ or $-20$ dB. It also takes on a very different shape for those high frequencies, particularly when the early echoes are added.

This ILD shrinkage with reverberation has been reported as the ratio of energy from the direct-path sound to the reverberant sound is decreased (Ihlefeld and Shinn-Cunningham, 2004). This effect is due to the generally isotropic nature of reverberation, as discussed in section 2.1.3. As more reverberant energy is added, either through a longer impulse response or a more energetic reverberant portion of the impulse response, the ILD of the direct-path will be mixed with the ILD of the reverberation. Because the amount of reverberation from a source is basically the same throughout a room, it will be similar at
both of a listener’s ears, regardless of the position of the source. Thus, it will have an ILD close to 0 dB as can be seen in figure 3.6(e). This also means that the ILD decreases with larger source-to-listener distances, as the direct-to-reverberant ratio decreases.

In figure 3.6(c), the ILD has many observations at the ILD of (b), but also has many observations close to 0 dB. This can be seen as a bimodality or a skew towards the origin. Because of the nonstationary and sometimes impulsive nature of speech, the direct-to-reverberant ratio of speech varies with time. When this ratio is high, the ILD cues are closer to the direct-path cues in (a), and when it is low, the ILD cues are closer to the late reverberation only cues in (e). In particular, at the beginning of utterances and after bursts preceded by short pauses, such as stop consonants, reverberation hasn’t had time to set in, and the ILD is closer to its value for the direct-path signal. This process is related to the precedence effect (Litovsky et al., 1999), and predicting a similar effect for ITD cues from monaural spectral features has been investigated by Wilson and Darrell (2006).

3.1.3.3 Additive noise

The IPD and ILD noise is illustrated in figure 3.7 for the anechoic signal at 75° mixed with additive noise. It was generated by convolving the signal with the direct-path portion of the BRIR and then adding speech-shaped noise to it at the same “direct-to-reverberant” ratio as the reverberation, approximately 10 dB. The speech shaped noise was filtered to match the average spectrum of 15 utterances chosen at random from the TIMIT corpus. These utterances were used in other experiments in this work as well, particularly section 3.2.3 and section 5.4.

The most notable feature of these plots is that the interaural parameters are very concentrated between 2.5 and 6 kHz and very diffuse outside of that range. This is most likely due to variation in per-frequency SNR cause by a mismatch between the energy profiles of this utterance and the speech shaped noise. The ILD plot is similar to figure 3.6(c), although the bimodality is more clearly defined for the additive noise while the early echoes in the reverberant case create more fine structure, i.e. rapid variation in the means of parameters with frequency. The IPD plot is similar to figure 3.5 as well, but similarly
lacks fine structure from the early echoes. Subfigure (b) shows the difference between the IPD and the mean of the direct path, as in figure 3.5(d)–(f). Notice that the mode across frequency is quite close to 0, meaning that IPD cues are preserved with additive noise.

3.1.4 W-Disjoint orthogonality

A property of speech that makes it easier to separate than stationary processes is windowed-disjoint orthogonality, abbreviated W-disjoint orthogonality or WDO (Yilmaz and Rickard, 2004). As a result of the “sparsity” of speech in the time-frequency domain, when two speech utterances are mixed together, one of them is generally much more energetic than the other at any given time-frequency point.

Figure 3.8 shows this W-disjoint orthogonality property for the example mixture under consideration in this chapter. Both plots are histograms of the difference in energy between the two sources at corresponding time-frequency points. The first is for anechoic sources, the second is for reverberant. The mixture of two sources at a particular point will be just the louder of the two, if the difference in energies is at least 6 dB (Roweis, 2001). The lines on the plot contain the region for which the two sources are within 6 dB of one another, and fewer than 25% of the points lie in this region. Thus it is generally a reasonable assumption that one source dominates the other at any particular time-frequency point.

Although speech has a wide bandwidth, the energy of an utterance is not spread evenly over the time-frequency plane. For example, the harmonicity of vowels means that most of the energy of a vowel is concentrated at the harmonics and very little in the spaces between harmonics. Similarly, for stop consonants, energy is concentrated in the burst, with very little during the stop. Other types of phones are sparse in the time and/or frequency domains to varying degrees. Fricatives, even though they are not sparse in high frequencies, have very little energy at low frequencies.
3. Statistics of interaural parameters

3.2 Ground truth separations

The Ideal Binary Mask, also known as the oracle binary mask, has been proposed as an upper bound on the performance of source separation algorithms that generate binary time-frequency masks (Li and Wang, 2009). It is “ideal” both in being optimal in terms of signal-to-distortion ratio and in being unattainable from real-world mixtures. This is because it is created using knowledge of the signals before they were mixed. A similar mask, which we call the ideal Wiener mask, or just Wiener mask, is a continuous mask that is optimal in terms of minimizing the mean squared error between the magnitude spectrogram of a signal and the reconstruction of its magnitude spectrogram from a mixture. It is also created using knowledge of the signals before they were mixed. In this section, we discuss oracle masks and propose a similar upper bound on algorithms that perform time-frequency masking using only point-wise interaural parameters. This section is based on Mandel and Ellis (2009).

This upper bound, which we call the Ideal Interaural Parameter Mask (IIPM), has access to the pre-mixed signals, but creates a binary time-frequency mask based solely on interaural level and phase differences. All points at a given frequency having a particular ILD and IPD must be either included or excluded from the mask together. By comparing the performance of such an estimator to that of the ideal binary mask, it is possible to determine the separation power of the interaural parameters in reverberation and additional separation performance that must be sought through other means, e.g., monaural source separation, source modeling, dereverberation, etc.

3.2.1 Oracle masks

All of the oracle masks described here are computed based on the energy in two signals, what we call the desirable energy and the undesirable energy. Desirable energy is the energy from sources that we would like to extract from a mixture, while undesirable energy is all of the other energy in the mixture. The optimal mask in the minimum mean squared error (MMSE) sense, i.e. for reconstructing the magnitude spectrogram of the desirable signal from the mixture, is the Wiener mask. Assuming that the desirable and undesirable signals are uncorrelated, this mask is approximately (Bodden, 1993)

$$M_W(\omega, t) \approx \frac{|S(\omega, t)|^2}{|S(\omega, t)|^2 + |N(\omega, t)|^2}$$  \hspace{1cm} (3.8)

where $S(\omega, t)$ is the STFT of the desirable signal and $N(\omega, t) = \sum N_i(\omega, t)$ is the STFT of the combined undesirable signals. The ideal binary mask (IBM) is constructed from the same signals according to

$$M_O(\omega, t) = \begin{cases} 1, & |S(\omega, t)| \geq |N(\omega, t)| \\ 0, & |S(\omega, t)| < |N(\omega, t)|. \end{cases}$$  \hspace{1cm} (3.9)

This mask can be considered to be a thresholded version of the Wiener filter mask (Ellis, 2006).
Table 3.1: The six oracle masks used in this work. + indicates that the mask treats a signal component as desirable, − indicates that the mask treats it as undesirable. Signal components are: target direct-path, target reverberation, masker reverberation, and masker direct-path. Note that early echoes are included in reverberation.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Target DP</th>
<th>Target Rev</th>
<th>Masker Rev</th>
<th>Masker DP</th>
</tr>
</thead>
<tbody>
<tr>
<td>DP-Oracle</td>
<td>Binary</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>DP-Wiener</td>
<td>Wiener</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Oracle</td>
<td>Binary</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>Wiener</td>
<td>Wiener</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
</tr>
<tr>
<td>OracleAllRev</td>
<td>Binary</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>WienerAllRev</td>
<td>Wiener</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
</tr>
</tbody>
</table>

Under certain conditions on hop size and window shape, the ideal binary mask is the optimal binary mask in terms of minimum mean squared error. This can be converted to a signal-to-noise ratio measured as

$$\text{SNR}_{Li} \equiv 10 \log_{10} \frac{\sum_t s^2(t)}{\sum_t (s(t) - \hat{s}(t))^2}$$  \hspace{1cm} (3.10)

where \(s(t)\) is the target signal and \(\hat{s}(t)\) is the estimate of that target signal separated from a mixture (Li and Wang, 2009). While this mask is optimal locally for each time-frequency point, and for each spectrogram frame, the overlapping of frames can lead to suboptimal global performance. In such cases, however, the performance of the ideal binary mask is generally close to optimal. Leaving those issues aside, we consider the oracle binary mask to be a close approximation to the optimal binary time-frequency mask.

One aspect of these masks that has not received much attention, but which we believe to be important, is the definition of the desirable signal. Because much of the initial work on the ideal binary mask (Roman et al., 2001; Roweis, 2001) was formulated for anechoic signals, its application to reverberant signals has followed closely in the same thread. Typically, this means that sources are convolved with an impulse response or reverberated in some way individually, and then combined to form the mixture. Signal \(s(t)\) or \(S(\omega, t)\) is then defined as the target source after it has been spatialized or reverberated.

We believe, however, that all reverberation, even reverberation from the target source, should be considered to be undesirable because it is detrimental to intelligibility (Lochner and Burger, 1964). We thus include the task of dereverberation in the task of source separation and use evaluation metrics that compare the output of source separators to the direct-path target signal only. Note, however, that it is impossible, with a binary mask, to recover the original, pre-spatialized target source. This is because the direct-path of the BRIR imparts a filtering due to the anechoic HRTF to the signal at each ear, as can be seen in comparing figure 2.2(d) and (a). Thus we focus on the goal of isolating the direct-path signal from the mixture instead of recovering the original unspatialized signal.

This distinction between the direct-path and the reverberation of the target and masker...
3. Statistics of interaural parameters

signals leads to the definition of six oracle masks. See section 6.4 for an evaluation metric based on this classification. While all of the ground truth masks are constructed based on the ratio of desirable energy to undesirable energy at each time-frequency point, these masks differ in their classification of these four types of energy as desirable or not. Table 3.1 indicates the desirability of each signal in each mask.

The DP-Oracle mask is a binary mask that only considers the target direct-path signal to be desirable. This mask provides an upper bound on masking separation performance under the metrics that we use, because the metrics also consider only the target direct-path signal to be desirable. It is used in experiments in chapters 5 and 6. The Oracle mask is the same, but also considers the reverberation of the target signal to be desirable. This is the mask that other authors have used and it provides an upper bound on separation performance under metrics that also consider target reverberation to be desirable. It is used in experiments in chapters 5 and 6. The OracleAllRev mask additionally considers the reverberation of the interfering signal to be desirable. This mask is proposed as an idealized comparison for the algorithmic separations described in chapter 6, which tend to successfully reject the direct-path signal from the interfering source, but have more difficulty rejecting its reverberation.

For each binary mask, there is a corresponding Wiener filter mask computed using equation (3.8) as the ratio of the desirable energy to the total energy at each time-frequency point. The DP-Oracle mask considers only the direct-path signal of the target source as desirable. The Wiener mask is the same, but also considers the reverberation from the target signal to be desirable. And finally, the WienerAllRev mask additionally considers the reverberation from the interfering source to be desirable. These masks are used in the automatic speech recognition experiments in chapter 6.

In constructing these masks, early echoes are considered to be part of the reverberation. In experiments with masks similar to those listed in table 3.1, but treating early echoes differently, performance was not noticeable different than for those listed in table 3.1. In particular, masks considering only direct-path and early echoes of the target to be desirable performed similarly to the DP masks. And masks considering only direct-path and early echoes of the masker to be undesirable performed similarly to the AllRev masks. This could be because the early echoes are 10 dB less energetic than the direct-path signal in these impulse responses (see figure 6.4). They also might not affect the masks much because they always occur in time-frequency points containing significant direct-path energy.

Note that because the observations are in stereo, an oracle mask could be produced for each ear. In practice, however, we do not do this because the systems in our evaluations only generate a single mask for both ears. We instead average the left and right signals together and build the oracle mask based on those averages.

3.2.2 Ideal Interaural-parameter mask

The Ideal Interaural Parameter Mask (IIPM) uses the ILD and IPD at each frequency to construct a mask similar to the ideal binary mask. From knowledge of the separated sources, the IIPM creates models for the interaural parameters weighted by the target and
masker energies. Specifically, it models them with kernel density estimators (KDEs) (Parzen, 1962), nonparametric models similar to histograms.

These kernel density estimates are created by weighting the observation of the training mixture at each time-frequency point by the energy of the signal at that point. Let $x(\omega, t) = [a(\omega, t) \phi(\omega, t)]^T$ be the a vector containing interaural level and phase differences of the mixture at frequency $\omega$ and time $t$. The KDEs describing the target and interference interaural parameters are, respectively,

$$k_T(\omega, x') = \sum_I w_I(\omega, t) \mathcal{N}(x' | x(\omega, t), \Sigma_I(\omega, t))$$

$$k_I(\omega, x') = \sum_I w_I(\omega, t) \mathcal{N}(x' | x(\omega, t), \Sigma_I(\omega, t))$$

where $x'$ is the arbitrary (ILD,IPD) point at which the estimate is being evaluated, $w_T$ and $w_I$ are the energy of the individual target and masker observations, and the covariance functions $\Sigma_T$ and $\Sigma_I$ are diagonal and are set using Silverman’s rule-of-thumb (Silverman, 1986, p. 48). Note that the density is estimated separately for each frequency channel. Note also that both KDEs are formed from the same points, the interaural parameters from the mixture, but weight those points differently using knowledge of the unmixed sources. We use the bounded error complexity reduction of Ihler (2005) to remove redundant kernels in the estimator while minimally distorting the modeled density. See Figure 3.9 for an example of target and masker KDEs and the decision boundary they induce at one particular frequency. The IIPM is then created according to

$$M_{IIPM}(\omega, t) = \begin{cases} 1, & \frac{k_T(\omega)}{k_I(\omega)} \geq \gamma \\ 0, & \text{otherwise} \end{cases}$$

where $\gamma$ is a user-defined threshold. As with the Ideal Binary Mask, we use $\gamma = 1$, meaning that a point in the interaural spectrogram is included if the KDE of the target energy at that pair of interaural parameters is larger than the KDE of the masker energy.

Two different IIPMs can be defined. The first separates signals using the interaural parameters calculated directly from each of the signals involved in a given mixture. We refer to this IIPM as the “training” IIPM. If the bandwidth of the kernels were infinitely

**Figure 3.9:** Example kernel density estimates in dB of (a) target energy, (b) masker energy, (c) energy ratio between target and masker with lines indicating the decision boundary at 0 dB. From the 4125 Hz band, target at 0° and masker at 90°.
3. Statistics of interaural parameters

(a) 438 Hz  
(b) 1547 Hz  
(c) 5141 Hz

Figure 3.10: Correlations between time-frequency points in the DP-oracle mask. Each plot is the correlation between a point at the indicated frequency and zero lag and its neighbors in time and frequency.

small, this IIPM would revert to the Ideal Binary Mask. This can be thought of as over-fitting the density estimator to the particular mixture under analysis, and would have a very low leave-one-out cross-validation likelihood. The training IIPM avoids this by setting the bandwidths according to the rule-of-thumb estimate, which tends to create favorable leave-one-out cross-validations.

The second IIPM uses interaural parameters calculated from different signals, but passed through the same impulse responses as the signals being separated. We refer to this IIPM as the “testing” IIPM. It still uses oracle knowledge of the unmixed signals to construct the KDEs, but to a lesser extent than the “training” IIPM does. It is guaranteed not to over-fit the test data. This IIPM is trained on a set of seven five-digit sentences from the target speaker, a total of 14 seconds of audio, which exhausted our memory resources. The threshold in each band implicitly includes an estimate of the signal-to-noise ratio, so if the per-band SNR between testing and training is significantly different, the threshold should be adjusted. We used similar utterances from the same speaker in the two conditions, so we did not adjust the threshold.

3.2.3 Correlations between mask points

One feature of these ground truth separation masks that should be useful in designing mask-based separation algorithms is the correlations between time-frequency points. When a point in a mask belongs to one source, it is likely that nearby points also belong to the same source. These correlations are not actually used in the systems described in subsequent chapters, but could be useful in future work.

This correlation can be seen in figure 3.10, which shows that nearby points’ mask values can have correlation coefficients as high as 0.8, with quite a few above 0.7. These plots were made by analyzing the DP-Oracle masks for 90 different mixtures of two sources in reverberation. The utterances used were 15 sentences from the TIMIT acoustic phonetic corpus (Garofolo et al., 1993). The DP-Oracle masks depend to a large extent on the sources and only to a small extent on the BRIRs used to mix them, so the fact that the 90 mixtures
only included 15 different sources might have increased the amount of correlation that would have been found between masks made from a more diverse set of signals.

These correlations were calculated by aggregating sufficient statistics over all of these masks. The masks were first rasterized, converting each region plotted in figure 3.10 into a long vector. These long vectors were advanced one frame at a time, so that adjacent vectors shared most of the same information, but in entries shifted relative to one another. The calculation then used the outer product of these vectors with the mask values at the three frequencies shown, along with the mean and standard deviation of each of the individual dimensions of both vectors.

A number of trends are visible in figure 3.10. First, for the mask point at 438 Hz, correlations are greatest at the same frequency. Correlations are also high at zero lag and at harmonically related frequencies. There is an interesting negative correlation between the mask at this frequency and the high frequencies at zero lag, possibly because many syllables are either dominated by low or high frequencies, but not both, e.g. vowels and fricatives, respectively.

The mask point at 1547 Hz is correlated with a relatively small group of points around it. These points are neighbors in frequency at small lags. Compared to the point at 438 Hz, these correlations are across more frequencies and fewer lags.

The mask point at 5141 Hz is correlated with many other points, especially other frequencies at zero lag. It is also negatively correlated with points at similar frequencies, but lags of \(~\sim\) 150 ms before and after it. This negative correlation could be a result of the 3–4 Hz temporal modulation rate of speech (Drullman et al., 1994a,b), if the two sources alternate at that rate. These oscillations are visible in the low frequency point’s correlation pattern as a positive correlation with high frequencies at a lag of \(~\sim\) 150 ms.

Note that all of these correlation patterns are approximately the same duration, except for the very specific temporal pattern of the lowest frequency. This is rather surprising, as one would expect the temporal correlation to decrease for higher frequencies if this were a result of bandwidths that were roughly proportional to center frequencies, as are observed in many mechanical resonance structures. The temporal modulation rate of speech could account for this regularity across frequency, however. Another possibility is that it is a result of the overlap of 75% between adjacent frames.

### 3.3 Summary

This chapter provided an in-depth introduction to the interaural parameters of sources passed through the binaural room impulse responses introduced in section 2.1. It discussed the interaural parameters that will be used in subsequent chapters for source localization and separation: the interaural phase, time, and level differences. It demonstrated the effects of each of the portions of the impulse response on each of these parameters. Specifically, the direct-path signal establishes the main trend in both IPD and ILD, which the early echoes disturb to some extent. Reverberation increases the variance of the IPD measurements and pushes the mean of the ILD measurements towards 0 dB. The chapter then discussed the convention that will be used in subsequent chapters of treating reverberation from
the target source as noise, and distinguished between six oracle masks, including the DP-Oracle and Oracle masks.
3.3 Summary