Chapter 4

Localization

This chapter, based on (Mandel and Ellis, 2006), discusses a number of approaches for localizing sound sources in azimuth, and generalizes them in two different ways. The first generalization, in section 4.1, expands the definition of cross-correlation from the time domain into the time-frequency domain, so that it can be calculated over arbitrarily shaped regions of a spectrogram. This top-down localization combines the information-gathering power of broad-band cross-correlation with the robustness to concurrent sources of bottom-up methods (e.g. Jourjine et al., 2000), which pool together narrowband localizations.

Section 4.2 discusses the von Mises circular probability distribution and generalizations of it for use in modeling the interaural phase difference at individual spectrogram points. It explains how these distributions can be used to create a probabilistic cross-correlation related to the more flexible time-frequency cross-correlation discussed in section 4.1. It also discusses recent results on the equivalence between the likelihood of a particular interaural time difference under a von Mises probability model and the generalized cross-correlation technique known as the Phase Transform (PHAT). Thus these generalizations of circular distributions are further generalizations of PHAT.

Finally, as in (Mandel and Ellis, 2006), section 4.3 presents experiments that show that this probabilistic method can localize sources using shorter observations than the generalized cross-correlation. The experiments also show that this model works well not just in reverberation, but also in additive speech-shaped noise with a similar signal-to-noise ratio and when localizing multiple sources in a mixture, validating the convolutional noise model for both types of noise.

This chapter focuses on the localization of sound sources using only the interaural phase difference, ignoring the interaural level difference. As discussed in section 2.3, humans predominantly use IPD to localize sources in azimuth, particularly at frequencies below 1.5 kHz (Wightman and Kistler, 1992), and barn owls have been shown to use it for frequencies up to 7 kHz (Moiseff and Konishi, 1981). The interaural level difference will be re-introduced in chapter 5.

4.1 The generalized cross-correlation

As discussed in section 3.1.2, when a single sound source, $x(t)$, is received at two microphones in different locations, the resulting observations, $\ell(t)$ and $r(t)$ are

$$\ell(t) = a_\ell x(t - \tau_\ell) * n_\ell(t) \qquad r(t) = a_r x(t - \tau_r) * n_r(t). \tag{4.1}$$

In this chapter we assume that a_ℓ, a_r, τ_ℓ, and τ_r are frequency independent, but we relax this assumption in chapter 5. In addition, additive noise could be present and there could be other filtering of the signals, but we assume that they are not present. Even with these assumptions, this model has been found to work well in practice and we explore its properties here. To the extent that the observations follow this model, the left and right channels are related to each other as

$$\ell(t) = a_{\ell r} r(t - \tau_{\ell r}) * n_{\ell r}(t) = a_{\ell r} r(t) * \delta(t - \tau_{\ell r}) * n_{\ell r}(t) \tag{4.2}$$

where the per-ear delays, gains, and noises have been combined into the relative qualities, $a_{\ell r}$, $\tau_{\ell r}$, and $n_{\ell r}(t)$.

The relative delay between the two signals can be estimated through their cross-correlation, defined as

$$r_{\ell r}(\tau) \equiv \sum_{t=0}^{N-1} \ell(t) r(t - \tau) \qquad \tau = -N + 1 \ldots N - 1 \tag{4.3}$$

which can be rewritten as a convolution

$$r_{\ell r}(\tau) = \sum_{t=0}^{N-1} \ell(t) r(-(\tau - t)) = \ell(t) * r(-t). \tag{4.4}$$

Under the model of equation (4.2), this becomes

$$\begin{aligned} r_{\ell r}(\tau) &= a_{\ell r} r(t) * \delta(t - \tau_{\ell r}) * n_{\ell r}(t) * r(-t) && (4.5)\\ &= a_{\ell r} r_{rr}(\tau) * \delta(\tau - \tau_{\ell r}) * n_{\ell r}(\tau) && (4.6) \end{aligned}$$

where $r_{rr}(\tau)$ is the auto-correlation of $r(t)$. Thus the cross-correlation is a copy of the auto-correlation shifted by the delay $\tau_{\ell r}$ and since the auto-correlation is peaked at 0, the cross-correlation is peaked at $\tau_{\ell r}$. Taking the Fourier transform of both sides of this equation gives

$$S_{\ell r}(e^{j\omega}) = L(e^{j\omega}) R^*(e^{j\omega}) = a_{\ell r} S_{rr}(e^{j\omega}) e^{-j\omega\tau_{\ell r}} N_{\ell r}(e^{j\omega}). \tag{4.7}$$

The convolution in time has become multiplication in frequency, the reversal in time has become conjugation, and the delay has been become a pure phase term that is linear in frequency.

The time-delay of arrival between the two signals, $\tau_{\ell r}$, can thus be estimated as

$$\hat{\tau} = \operatorname*{argmax}_\tau r_{\ell r}(\tau). \tag{4.8}$$

By performing this calculation in the frequency domain and then taking the inverse transform, the it becomes

$$r_{\ell r}(\tau) \propto \sum_{k=-N/2}^{N/2-1} L(e^{j\omega})R^*(e^{j\omega})e^{j\omega\tau} \qquad \omega = \tfrac{2\pi k}{N} \tag{4.9}$$

where the proportionality comes from the $\frac{1}{N}$ in the inverse Fourier transform, which will be ignored from now on. For notational convenience, subsequent equations will elide k and ω and indicate a sum over ω. Because the argmax of $r_{\ell r}(\tau)$ is used in estimating the delay, the values of that function only matter relative to one another, not on an absolute scale, i.e. any monotonic warping can be applied to it without affecting the estimate.

In this form, the cross-correlation can be evaluated at non-integral values of τ. Because the $r_{\ell r}(\tau)$ signal is band-limited, however, evaluating equation (4.9) at such values merely interpolates between the integral values. Care must be taken when using the discrete Fourier transform to compute this sum in this situation, however. The $e^{j\omega\tau}$ term must be computed for $k = -\frac{N}{2} \ldots \frac{N}{2} - 1$ and not for $k = 0 \ldots N-1$. For integral values of τ these two ranges of k will produce the same $e^{j\omega\tau}$ vector, but for non-integral values of τ, the version using positive ks introduces a Hilbert transform-like phase offset that is constant with frequency.

The Generalized Cross-Correlation (GCC) (Knapp and Carter, 1976) introduces a weighting function into equation (4.9)

$$g(\tau) = \sum_{\omega} \psi(e^{j\omega})L(e^{j\omega})R^*(e^{j\omega})e^{j\omega\tau}. \tag{4.10}$$

Various algorithms in this family can be defined by using different $\psi(e^{j\omega})$ functions, some of which depend on the specific observations $L(e^{j\omega})$ and $R(e^{j\omega})$, and others on their statistical properties or estimates thereof. Optimal localization results are possible in this framework when the signals and noise fit certain models. Other weighting functions work well empirically in many situations, one of which is the Phase Transform (PHAT), in which the weighting factor $\psi(e^{j\omega})$ cancels the magnitudes of the left and right signals,

$$\psi_{\text{PHAT}}(e^{j\omega}) = \frac{1}{|L(e^{j\omega})||R^*(e^{j\omega})|} \tag{4.11}$$

so that the GCC cost function becomes

$$g(\tau) = \sum_{\omega} \frac{L(e^{j\omega})}{|L(e^{j\omega})|} \frac{R^*(e^{j\omega})}{|R^*(e^{j\omega})|} e^{j\omega\tau} \tag{4.12}$$

This whitening works well for broadband signals, but amplifies background noise when the target signal is in fact narrowband. In this case, another $\psi(e^{j\omega})$ can be used. If the signal embedded in $L(e^{j\omega})$ and $R(e^{j\omega})$ is limited to the frequency band ω_{min} to ω_{max}, then using values outside of this region would add to the noise and would not help the estimate of the delay between the two channels. In this case, the following weighting function could

be used

$$\psi_{\mathrm{BL}}(e^{j\omega}) = \begin{cases} 1 & \omega_{\min} \leq \omega \leq \omega_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{4.13}$$

In effect, this changes the set of ωs over which the sum is taken in equation (4.9). This band-limited weighting could also be combined with PHAT if the magnitude of the target signal varies significantly within this frequency band.

4.1.1 Comparison to other approaches

When calculated explicitly in the frequency domain, equation (4.9) can be considered to be a top-down localization method. This means that it can be calculated for a single value of τ or for any number of values. This can be thought of as testing hypotheses against the observations. The problem with using a regular cross-correlation, however, is that it tends to become unreliable in the presence of multiple sources. Bottom-up localization methods, on the other hand, include DUET (Jourjine et al., 2000), and the straightness-based measure of (Stern and Trahiotis, 1995), can successfully localize multiple simultaneous sources by performing some form of localization separately at each frequency band and then attempting to assemble the local estimates into global estimates. Each of the component narrowband localizations, however, suffers from a multimodal cross-correlation.

The probabilistic approach, described later in this chapter and based on the frequency-restricted GCC weighting of equation (4.13), combines the best aspects of these two methods. It compares top-down hypotheses using narrowband features. It provides a structured way to combine information across frequency bands. Another such method was proposed by Liu et al. (2000), although it operates on ITDs instead of IPDs. Because delays affect IPD linearly with frequency, however, we find that IPD provides a more useful representation for localization.

The frequency-domain GCC framework can be used to compute cross-correlations in auditory filterbanks much more quickly than when processing them in the time domain (e.g. Faller and Merimaa, 2004). The time domain calculation begins by filtering both channels through a set of bandpass filters with impulse responses $\{h_f(t)\}$, where f is the center frequency of each channel. This creates a number of bandpassed versions of each channel that are sampled at the original signal's sampling rate, $\{\ell(t) * h_f(t), r(t) * h_f(t)\}$. Cross-correlations are then performed between the two channels in corresponding frequency bands. In the convolution notation of equation (4.4), this can be written

$$r^f_{\ell r}(\tau) = \ell(t) * h_f(t) * r(-t) * h_f(-t). \tag{4.14}$$

It can also be computed recursively using a running cross-correlation (e.g. Faller and Merimaa, 2004).

In order to speed up this computation slightly, each per-band cross-correlations can be performed in the frequency domain using equation (4.9)

$$r^f_{\ell r}(\tau) = \sum_\omega L_f(e^{j\omega}) R^*_f(e^{j\omega}) e^{j\omega\tau} \tag{4.15}$$

$$= \sum_\omega L(e^{j\omega}) H_f(e^{j\omega}) R^*(e^{j\omega}) H^*_f(e^{j\omega}) e^{j\omega\tau} \tag{4.16}$$

$$= \sum_\omega |H_f(e^{j\omega})|^2 L(e^{j\omega}) R^*(e^{j\omega}) e^{j\omega\tau} \tag{4.17}$$

where $H_f(e^{j\omega})$ is the discrete Fourier transform of $h_f(t)$. This requires two Fourier transforms for each band. Note, however, that this equation is a generalized cross-correlation, with $\psi_f(e^{j\omega}) = |H_f(e^{j\omega})|^2$ and the dependence on the bandpass filter is entirely contained in the weighting function. This means that the Fourier transform can be computed just once on the original signals and different GCC weightings used to generate the cross-spectra of all of the bandpass signals affording significant computational savings. Each band still requires an inverse Fourier transform to convert the cross-spectrum to the cross-correlation, but in certain circumstances it is possible to avoid this computation. Larger speedups can also be achieved if the filters $H_f(e^{j\omega})$ are nonzero at a small number of frequencies or if they are a constant shape, and only vary in center frequency.

4.1.2 Generalization across time

Similarly to its generalization in frequency, the cross-correlation can be generalized to multiple time points in a time-frequency representation like the short-time Fourier transfer (STFT). Define the GCC at time t as

$$g_t(\tau) = \sum_\omega \psi_t(e^{j\omega}) L(\omega, t) R^*(\omega, t) e^{j\omega\tau} \tag{4.18}$$

where $L(\omega, t)$ and $R(\omega, t)$ are the STFTs of the left and right signals, respectively. There are two ways to combine estimates from multiple instants of time. Either the $g_t(\tau)$ functions can be added together before taking the argmax in equation (4.8), or the argmax can be taken first and a histogram of $\hat{\tau}_t$ values can be accumulated. The technique in which multiple argmaxes are collated is called PHAT-Histogram by Aarabi (2002). The technique in which the argmax is applied to the sum of the cross-correlations fits into the framework of equation (4.9) and under it that equation can be rewritten

$$g(\tau) = \sum_{t=t_{\min}}^{t_{\max}} \sum_\omega \psi_t(e^{j\omega}) L(\omega, t) R^*(\omega, t) e^{j\omega\tau}. \tag{4.19}$$

Care must be taken to avoid over-counting observations, however. When the STFT is thought of as a filterbank, if the bands are sampled more frequently than their critical rate, correlated values will be combined, possibly biasing the estimates.

Using this formulation, the GCC, which began with a sum over all values of ω in a single frame of the STFT, can be taken over fewer frequencies or over more time frames. In fact, the sum can be taken over any arbitrarily shaped set of points in the STFT. It can even

be computed over a single point. If Ω is the set of time-frequency points over which the cross-correlation is taken, then the generalized generalized cross-correlation can be written simply as

$$g(\Omega;\tau) = \sum_{(\omega,t)\in\Omega} \psi(\omega,t)L(\omega,t)R^*(\omega,t)e^{j\omega\tau} \tag{4.20}$$

where $\psi_t(\omega)$ has been rewritten as $\psi(\omega,t)$ to emphasize the fact that it has into a two-dimensional window function, also known as a kernel.

When a different region is associated with each point in the spectrogram, a cross correlation can be assigned to each point based on this region,

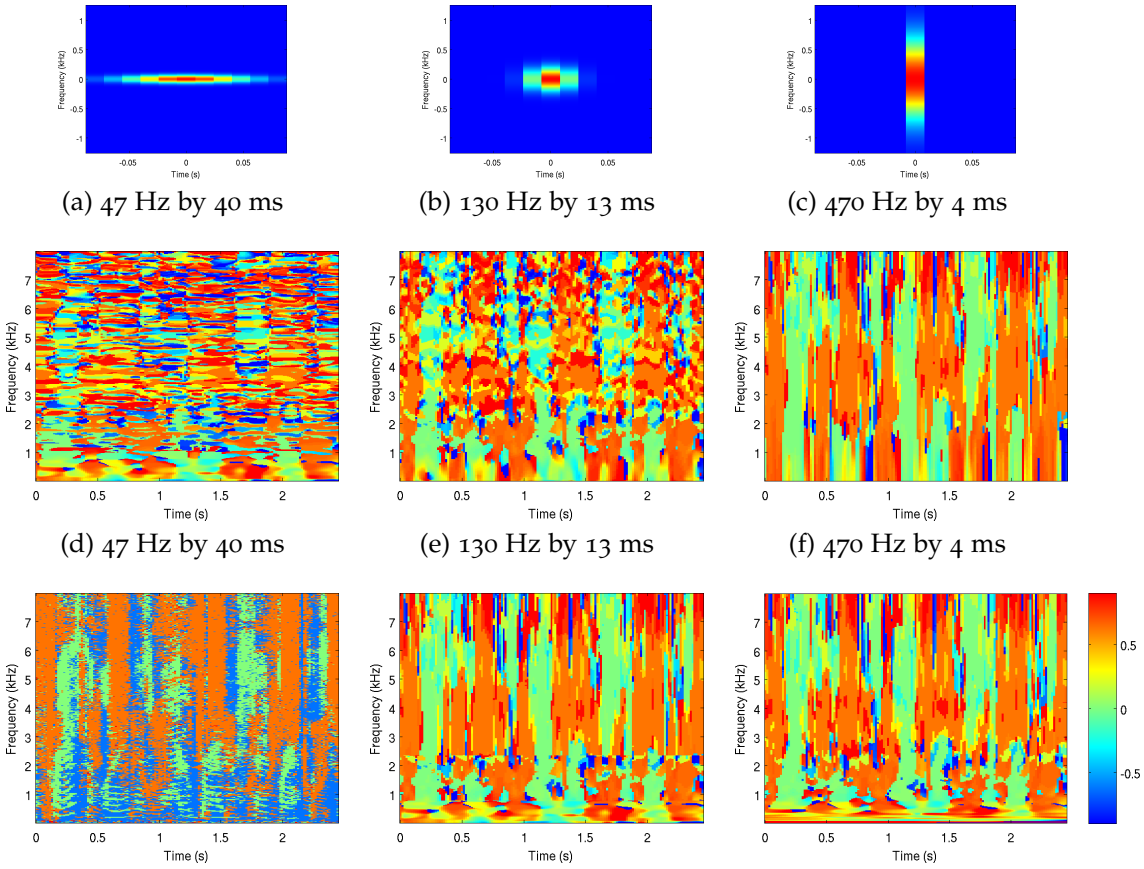
$$g(\omega,t;\tau) = \sum_{(\omega',t')\in\Omega(\omega,t)} \psi(\omega'-\omega,t'-t)L(\omega',t')R^*(\omega',t')e^{j\omega'\tau}. \tag{4.21}$$

where $\Omega(\omega,t)$ is the set of time-frequency points neighboring the point (ω,t). See figure 4.1 for examples of such a cross-correlation computed with four different kernels. The signal being analyzed is the example mixture from chapter 3, with two speakers in a reverberant environment, one located at 0° and one at 75°. Equation (4.21) is calculated for the 61 values of τ between -15 and 15 in steps of 0.5 (corresponding to -940 to 940 ms in steps of 30 ms) and then the argmax of equation (4.8) is taken over τ.

Figure 4.1(d)–(f) shows the cross-correlations that result from the use of the three Gaussian kernels shown in (a)–(c). Greater amounts of smoothing in either direction increases the reliability of the cross-correlation in regions of the spectrogram that are dominated by one of the sources. These estimates are less reliable at boundaries between such regions, because they are forced to combine conflicting information. In particular, as described in section 3.2.3 at high frequencies, a single source tends to dominate adjacent frequencies, but not necessarily adjacent times. Similarly, at low frequencies, a single source tends to dominate adjacent times, but not adjacent frequencies. Thus the kernel used in these regions should reflect these trends.

The ground truth plot in (g) was created using the DP-Oracle masks for the two sources, which are disjoint. Each point was assigned to the delay of the DP-Oracle mask that is active at it. If no mask was active at a particular point, it was colored blue to distinguish it from the two true sources. In the other plots, these areas should be randomly colored because the diffuse nature of the reverberation leads to random cross-correlation peaks. Designating this plot ground truth assumes that the direct-path signals are perfectly localized while reverberation is perfectly diffuse, containing no localization information. This is clearly not the case, but it is a useful simplifying assumption.

The kernel in (a) is the most "horizontal" of the three, with standard deviations of 47 Hz in frequency and 40 ms in time. Notice that its results, in (d), are correlated across time, but only slightly across frequency. This kernel agrees with the ground truth best at low frequencies, because at these frequencies, a single source tends to dominate adjacent points in time, but not in frequency. The kernel in (b) is more "symmetric" in time and frequency, having standard deviations of 130 Hz and 13 ms. Its results, in (e), are correlated moderately across time and frequency and it agrees with the ground truth best at mid frequencies. And the kernel in (c) is more "vertical" in time and frequency, having standard



(a) 47 Hz by 40 ms (b) 130 Hz by 13 ms (c) 470 Hz by 4 ms

(d) 47 Hz by 40 ms (e) 130 Hz by 13 ms (f) 470 Hz by 4 ms

(g) Approx. gnd. truth (h) Piecewise freq.-dep. (i) True freq.-dep.

Figure 4.1: Example of cross-correlation taken over axis-aligned Gaussian regions around each time-frequency point in a spectrogram. Values are the argmax across τ of the cross-correlation function. The signal being analyzed is the example mixture from chapter 3. The Gaussian kernels in (a)–(c) with the indicated standard deviations are used in (d)–(f). (g) DP-Oracle mask points assigned to appropriate locations with blue indicating no direct-path source present, (h) Combination of the three kernels at low, mid, and high frequencies, (i) True frequency-dependent kernel keeping the time-frequency product constant.

deviations of 470 Hz and 4 ms. Its results, in (f), are correlated across frequency and it agrees with the ground truth best for high frequencies.

These plots are combined in (h), using (d) for frequencies below 800 Hz, (e) for frequencies between 800 and 2400 Hz, and (f) for frequencies above 2400 Hz. Thus each of these kernels is used in the region where it best matches the ground truth. Taking this idea to its logical conclusion, in (i) a different kernel is used for every frequency band in the spectrogram

while maintaining a constant time-frequency product. This approach achieves greater accuracy at the expense of added computational complexity. Intermediate strategies could also be used that combine more than three, but fewer than 511 kernels, to select a particular accuracy-complexity trade off. Note that (h) resembles (i), except close to the boundaries between the three frequency bands where there are discontinuities. These discontinuities are mitigated as the number of bands increases. The time-frequency product for all of the kernels used in this section was selected to maximize the separation performance of the KernelXcorr algorithm used in the multi-source localization experiments in section 4.3.2.

4.1.3 Interaural coherence

A related quantity to the cross-correlation is the coherence between two signals (Faller and Merimaa, 2004; Allen et al., 1977; Carter, 1993). Whereas the cross-correlation is the argmax of $g(\tau)$ over τ, the coherence is the actual maximum value of the cross-correlation. In order for this value to be meaningful, however, the cross-correlation has to be *normalized* so that it is always within a fixed range. The normalized cross-correlation, which is closely related to $g(\tau)$, is defined as

$$\gamma_{\ell r}(\tau) = \frac{\sum_t \ell(t) r(t-\tau)}{\sqrt{\left(\sum_t \ell^2(t)\right)\left(\sum_t r^2(t)\right)}} = \frac{r_{\ell r}(\tau)}{\sqrt{r_{\ell\ell}(0) r_{rr}(0)}} \tag{4.22}$$

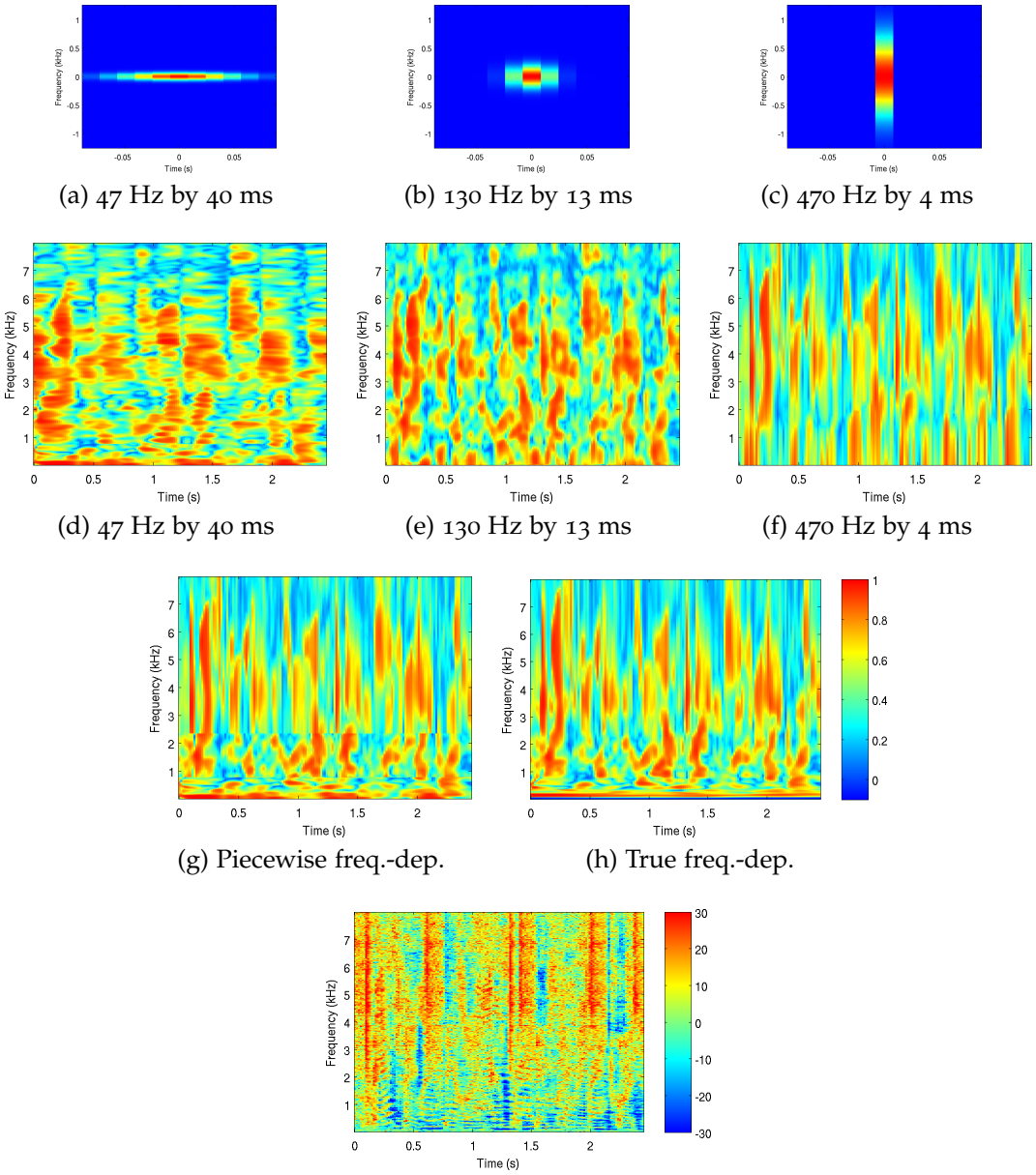
In the frequency domain, this becomes an inner product between unit-power versions of $L(e^{j\omega})$ and $R(e^{j\omega})$, and the Cauchy-Schwarz inequality ensures that this will always be between -1 and 1. Because the delay does not change the magnitude of any of the terms involved in the inner product, it does not affect this bound. Note that the savings of computing the filterbank cross-correlations as in equation (4.17) also apply to coherence computations.

This normalized cross-correlation is made up of a cross-correlations and two auto-correlations, each of which can be generalized in the same way that the cost function in equation (4.9) is generalized to cover arbitrary regions of the spectrogram while still maintaining the normalization of equation (4.22)

$$\gamma_{\ell r}(\Omega;\tau) = \frac{\sum_{(\omega,t)\in\Omega} \psi(\omega,t) L(\omega,t) R^*(\omega,t) e^{j\omega\tau}}{\sqrt{\left(\sum_{(\omega,t)\in\Omega} \psi(\omega,t)|L(\omega,t)|^2\right)\left(\sum_{(\omega,t)\in\Omega} \psi(\omega,t)|R(\omega,t)|^2\right)}}. \tag{4.23}$$

This array can be computed efficiently by precomputing certain quantities. First, the two magnitude spectrograms, $|L(\omega,t)|^2$ and $|R(\omega,t)|^2$, can be computed. Then, for each τ under consideration, the spectrogram $L(\omega,t)R(\omega,t)e^{j\omega\tau}$ can be computed. After that, the computation of $\gamma_{\ell r}(\Omega;\tau)$ is a simple matter of multiplying these quantities by the kernel $\psi(\Omega)$ and then summing.

See figure 4.2 for examples of such a coherence calculation, computed with three different kernels. This figure is complementary to figure 4.1 in that it contains the max, while figure 4.1 contains the argmax. This figure also uses the same kernels as figure 4.1. Note that figure 4.2 seems much less noisy than figure 4.1, even for computations with the same



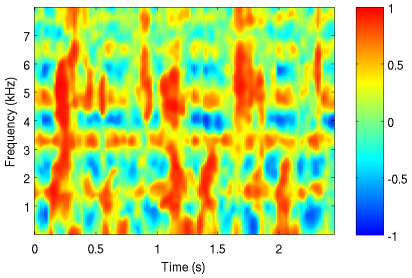
(a) 47 Hz by 40 ms (b) 130 Hz by 13 ms (c) 470 Hz by 4 ms

(d) 47 Hz by 40 ms (e) 130 Hz by 13 ms (f) 470 Hz by 4 ms

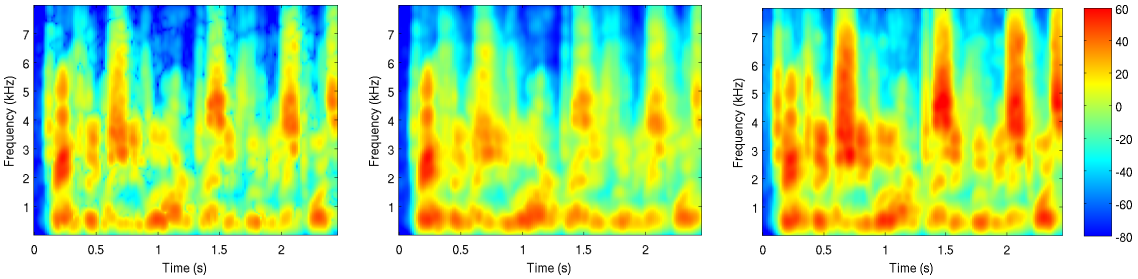
(g) Piecewise freq.-dep. (h) True freq.-dep.

(i) Direct-to-reverberant ratio

Figure 4.2: Example of coherence taken over axis-aligned Gaussian regions around each time-frequency point in a spectrogram. The signal is the example from chapter 3. The Gaussian kernels in (a)–(c) with the indicated standard deviations are used in (d)–(f). (g) Combination of the three kernels at low, mid, and high frequencies, (h) True frequency-dependent kernel keeping the time-frequency product constant, (i) Ratio of direct-path energy to reverberant energy, in dB.



(a) Normalized cross-corr. at $\tau = 0$



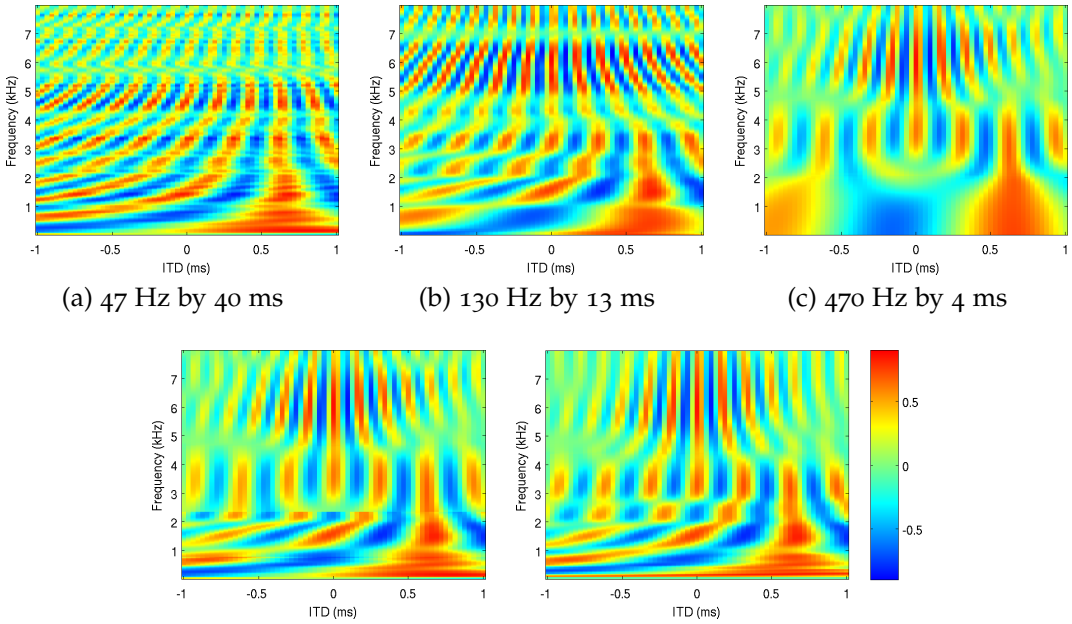
(b) Smoothed cross-corr. (c) Smoothed $|L(\omega, t)|^2$ (d) Smoothed $|R(\omega, t)|^2$

Figure 4.3: The three different spectrograms involved in the calculation of the normalized cross-correlation at $\tau = 0$ for figure 4.2(e).

kernel. This is because of the differences in the noise characteristics of the max versus the argmax when applied to a multimodal function. These values are multimodal in τ, as can be seen in figure 4.4. Specifically, the max is only perturbed slightly if the function jumps between modes, whereas the argmax displays a large discontinuity.

Similarly to figure 4.1, (d) is more smoothed across time and works better for low frequencies, (e) is more smoothed in general, and (f) is more smoothed across frequency and works better for high frequencies. The combination of the three plots in (h) also closely approximately the frequency-dependent calculation in (i), except at the edges of the frequency bands. The direct-to-reverberant energy ratio is shown in (g) as a proxy for real ground truth. Because the direct-path signal is supposed to be compact and the reverberation diffuse, time-frequency points with more direct-path than reverberant energy should be more coherent than those in which reverberation dominates. Thus the direct-to-reverberant ratio should be qualitatively similar to the coherence, and comparison with the coherence plots shows this to be the case. One notable difference is that the direct-to-reverberant ratio is generally higher at higher frequencies, which is not the case for the coherence.

The computation of the coherence in equation (4.23) is similar to that of (Allen et al., 1977), which was proposed for dereverberation of binaural signals. By performing calculations without the $e^{j\omega\tau}$ term, they were able to use just a single running cross-correlation instead of one for each value of τ required by equation (4.23). This τ dependence can be ignored when the smoothing kernel is only applied across time, but it is required to pool across frequency. Note that without any pooling, the coherence calculation yields 1 uniformly,



(a) 47 Hz by 40 ms (b) 130 Hz by 13 ms (c) 470 Hz by 4 ms

(d) Piecewise freq.-dep. (e) True freq.-dep.

Figure 4.4: Example of the multimodality of cross-correlation in τ, using axis-aligned Gaussian regions around each time-frequency point in a spectrogram. The signal being analyzed is the example from chapter 3, the time slice occurs at 0.96 s, and the sources are located at ITDs 0 and 0.7 ms. Gaussian kernel standard deviations are indicted.

as the cross-spectral magnitude at a single time-frequency point is the product of the magnitudes at each ear at that point. As shown in section 3.2.3 and figure 4.2, pooling across frequency is very important for capturing detail in the coherence at high frequencies. Pooling mainly across time will result in coherence plots more like figure 4.2(d), although using an auditory spectrogram will help bring it closer to (h), assuming that the time-scale of the smoothing is adjusted accordingly. Note that while Allen et al. (1977) reported that their algorithm subjectively decreased reverberation, Bloom and Cain (1982) found that it did not improve intelligibility.

4.2 Probabilistic cross-correlation

Each STFT-like "slice" of equation (4.21) for a given value of τ can be thought of as accumulating energy for that τ in $g(\tau)$. The τ with the largest $g(\tau)$ then "wins" the cross-correlation. By converting this accumulation of energy into an accumulation of probability or likelihood, we can take advantage of a number of nice features of probability, building the foundation for the source separation system described in chapter 5. After describing

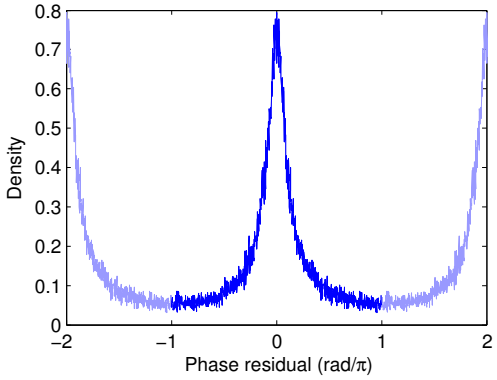


Figure 4.5: A histogram of measured IPD residual angles in a reverberant environment. Shown for two cycles of IPD.

the data that are being modeled and a family of probability distributions that does a good job of modeling these data, we will describe the probabilistic computation using them.

In section 3.1.3.1, we showed examples of the distribution of IPD angles around $-\omega\tau$. In general, we found that after factoring out this linear frequency dependence, the observations generally followed the same frequency-independent distribution. Thus we make the assumption that all observations in the time-frequency plane are independent and identically distributed (IID) conditioned on their true delay.

More formally, we now describe the model for these measurements. For any single point in time and frequency, define the observed IPD and the IPD residual as, respectively,

$$\phi(\omega,t) = \angle\frac{L(\omega,t)}{R(\omega,t)} \qquad \hat{\phi}(\omega,t;\tau) = \angle\left(\frac{L(\omega,t)}{R(\omega,t)}e^{-j\omega\tau}\right). \tag{4.24}$$

Without using the phase residual, phase circularity becomes a problem when observations originate from delays that are far from 0. Using the phase residual, this problem is shifted to delays that are far from the true delay. See figure 4.5 for a histogram of these residual values, $\hat{\phi}$, pooled over all of the time-frequency points in a spectrogram. Note that this histogram is peaked at 0 and periodic with a period of 2π, meaning that the histogram of the phase itself is peaked at $-\omega\tau$ and periodic with a period of 2π.

In subsequent subsections we describe various distributions with which this histogram can be modeled. Specifically, our goal is to identify a distribution that allows us to assign a likelihood to phase residual observations at individual time-frequency points.

4.2.1 The von Mises distribution

The von Mises distribution is a probability distribution defined over angles (Mardia and Jupp, 1999)

$$p(\theta \mid \mu_\theta, \kappa) = \mathrm{vM}_\theta(\theta; \mu_\theta, \kappa) = (2\pi I_0(\kappa))^{-1} \exp\{\kappa \cos(\theta - \mu_\theta)\} \tag{4.25}$$

where $\theta \in [-\pi, \pi)$ is the angle, $\mu_\theta \in [-\pi, \pi)$ is the mean angle of the distribution, κ is the concentration parameter, and $I_0(\kappa)$ is the zero-th order modified Bessel function of the first kind. The von Mises distribution can also be parametrized in a vector form, instead of the angular form

$$p(x \mid \mu_x, \kappa) = \mathrm{vM}(x; \mu_x, \kappa) = (2\pi I_0(\kappa))^{-1} \exp\{\kappa \mu_x^T x\} \tag{4.26}$$

where μ_x and x are unit vectors with μ_x pointing in the mean direction of the distribution on the unit circle. This vector parametrization can be generalized beyond two dimensions to hyperspheres of arbitrary dimension.

The von Mises distribution has many properties in common with the Gaussian distribution (Mardia, 1975; Gatto and Jammalamadaka, 2007). Both are maximum entropy distributions given the two lowest-order expectation constraints on their respective supports. Both are symmetric and have their mode and median at the mean. Both are in the exponential family and are self-conjugate, meaning that the conjugate prior for the mean parameter of each is the same distribution. Thus it is straightforward to compute the maximum likelihood or maximum *a posteriori* mean parameter of the von Mises distribution, μ_x. Although there is no closed form maximum likelihood estimate of the concentration parameter, κ, it can be computed numerically.

The likelihood and posterior calculations are as follows (Mardia and El-Atoum, 1976). Let x_i be a set of observations on the unit circle, with $i = 1 \ldots N$. These observations are assumed to be independent and identically distributed, so their likelihood is

$$p(x \mid \mu_x, \kappa) = \prod_i \mathrm{vM}(x_i; \mu_x, \kappa) \tag{4.27}$$

$$= \prod_i (2\pi I_0(\kappa))^{-1} \exp\{\kappa \mu_x^T x_i\} \tag{4.28}$$

$$= (2\pi I_0(\kappa))^{-N} \exp\Big\{\kappa \mu_x^T \sum_i x_i\Big\}. \tag{4.29}$$

If we assume that the mean μ_x is also distributed according to a von Mises distribution, with mean m and concentration k, then the prior on it can be written

$$p(\mu_x \mid m, k) = \mathrm{vM}(\mu_x; m, k) = (2\pi I_0(k))^{-1} \exp\{k m^T \mu_x\}. \tag{4.30}$$

The posterior distribution over μ_x having seen the observations x_i is then

$$p(\mu_x \mid x, \kappa, m, k) \propto \exp\Big\{\kappa \mu_x^T \sum_i x_i + k \mu_x^T m\Big\} \propto \mathrm{vM}(\mu_x; m', k') \tag{4.31}$$

where $k' = \kappa \| \sum_i x_i + \frac{k}{\kappa} m \|$ and $m' = \frac{\kappa}{k'} (\sum_i x_i + \frac{k}{\kappa} m)$.

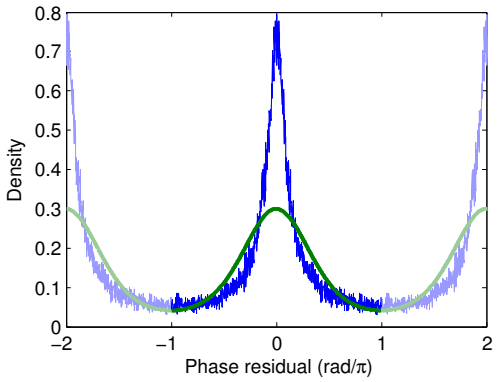


Figure 4.6: The histogram of figure 4.5 plotted with a von Mises distribution with matching moments. Shown for two cycles of IPD.

Figure 4.6 shows the same histogram of IPD residuals as figure 4.5, but superimposed is the likelihood of the von Mises distribution that best fits that data. Notice that the mode of the von Mises distribution is much wider than the mode of the histogram so that it can properly account for the probability mass in the angles diametrically opposite to it.

4.2.2 Maximum entropy circular distributions

Unfortunately, the von Mises distribution does not fit the histogram very well, in particular, the "tails" of our data are too heavy. Figure 4.6 shows that while the histogram is unimodal with a rather sharp peak, it still has significant probability density at all IPD residuals, even at π. Thus a von Mises distribution that matches the data in terms of the width of the main lobe gives too little probability to points far from the mean direction. Similarly, the von Mises distribution that matches the observations in terms of the concentration parameter gives too little probability to points close to the mode, and has a main lobe that is much too broad.

Thus in this section we examine a generalization of the von Mises distribution described by Mardia (1975) that can satisfactorily model both the width of the main lobe and the somewhat uniform background probability density at all angles. It would be possible to model our data with a mixture of a von Mises distribution and a uniform distribution, which is a von Mises distribution with $\kappa = 0$. We prefer, however, to take another approach based on maximum entropy circular distributions because of its elegance.

Just as distributions over the real numbers have moments, so do distributions over angles. These moments are called circular moments, and the kth circular moment of $p(\theta)$ is defined as (Gatto and Jammalamadaka, 2007)

$$M_k\{p(\theta)\} = \int_0^{2\pi} p(\theta) e^{jk\theta}\, d\theta = \int_0^{2\pi} p(\theta)\cos(k\theta)\, d\theta + j\int_0^{2\pi} p(\theta)\sin(k\theta)\, d\theta. \tag{4.32}$$

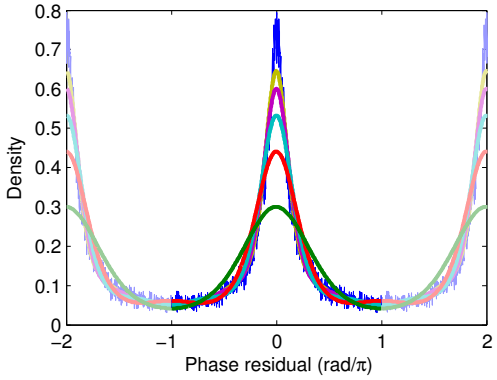


Figure 4.7: The histogram of figure 4.5 plotted with successive maximum entropy approximations to it. Shown for two cycles of IPD.

Note that this is a complex quantity, so that a single moment provides two constraints on the distribution.

Cover and Thomas (2006, Chap. 12) describe maximum entropy distributions, which are probability distributions that meet certain expectation constraints, but make as few other assumptions as possible. In particular, subject to the constraints that

$$E\{f_k(x)\} = \lambda_k = \int_{x \in S} p(x) f_k(x)\, dx, \tag{4.33}$$

the maximum entropy distribution over the interval $S \subseteq \mathbb{R}^n$ is

$$p(x) = c \exp\Big(\sum_k \alpha_k f_k(x)\Big) \tag{4.34}$$

where the α_k and c parameters are chosen to match the expectation constraints. A distribution satisfying any particular set of expectation constraints is not guaranteed to exist, but if it does, it is of this form and is unique.

If the constraint functions $f_k(\theta)$ are the first m circular moments, and the interval S is $[-\pi, \pi)$, then the maximum entropy circular distribution becomes (Mardia, 1975)

$$p(\theta) = c \exp\Big(\sum_{k=1}^{m} \alpha_k e^{jk\theta}\Big) = c \exp\Big(\sum_{k=1}^{m} \Re\{\alpha_k\} cos(k\theta) + j\Im\{\alpha_k\} \sin(k\theta)\Big). \tag{4.35}$$

In order to enforce symmetry of the IPD residual about the origin, we can use an even more constrained maximum entropy distribution in which all of the α_k are pure real, so that only the cosine term remains. See figure 4.7 for the histogram of IPD residuals with successive maximum entropy circular distributions superimposed. Note that these distributions can successfully capture the shape of the main peak while still giving sufficient probability to the "tails" of the distribution.

The von Mises distribution is the first order maximum entropy circular distribution. Specifically, the von Mises distribution's parameters, κ and μ_θ solve

$$\frac{I_1(\kappa)}{I_0(\kappa)} \cos \mu_\theta = \Re\{\alpha_1\} \qquad \frac{I_1(\kappa)}{I_0(\kappa)} \sin \mu_\theta = \Im\{\alpha_1\}. \tag{4.36}$$

Gatto and Jammalamadaka (2007) discuss the second order maximum entropy circular distribution, which uses the second circular moment as well.

A simple way to approximate this distribution, when there are many points to model is to match the moments of a circular histogram instead of computing each moment directly from the data. If $f(\theta)$ is the piecewise constant histogram function, then its kth circular moment is as described in equation (4.32). Because $f(\theta)$ is a histogram over angles, it is strictly periodic in 2π, its Fourier series is computed as

$$\lambda_k = \int_0^{2\pi} e^{jk\theta} f(\theta)\, d\theta \tag{4.37}$$

which is exactly the same as the computation of the moments in equation (4.32). Thus the maximum entropy distribution that matches the first m moments of the observed data is the truncated Fourier series approximation to the angular histogram of the observed data. Also because of the strict periodicity in 2π, the Fourier series is exactly equal to the discrete Fourier transform of the histogram, making it easy to compute numerically.

4.2.3 Top-down cross-correlation

A top-down cross-correlation, as described in section 4.1.1, requires a means of combining likelihoods of multiple time-frequency points. Such a calculation is particularly easy with the IPD residual noise because it is independent and identically distributed for all points in the spectrogram conditioned on the true delay. Specifically, the joint probability of observations at multiple points decomposes into a product of their marginal probabilities. The likelihood of a particular τ having observed the IPD in a region Ω is

$$p(\phi(\Omega) \mid \tau) = \prod_{(\omega,t)\in\Omega} p\left(\hat{\phi}(\omega, t; \tau)\right). \tag{4.38}$$

This likelihood can be used to compare the degree to which different values of τ fit the observed IPD residuals. Because it uses IPD residuals instead of IPDs, the points are all IID and zero-mean regardless of frequency, namely 0. This likelihood can serve as a probabilistic replacement for $g(\tau)$ from equation (4.21).

While equation (4.21) can be seen as a top-down cross-correlation, equation (4.38) is a better reasoned and more robust way of performing such a computation. As will be shown in chapter 5, the probabilistic approach also lends itself to use in separation systems and combination with other probabilistic cues and separators. The probabilistic approach can also be applied to a single time-frequency point, which is problematic for the direct cross-correlation methods. Both approaches are top-down because instead of converting IPD to ITD directly, each ITD hypothesis is tested against a set of IPD observations. See figure 4.8 for a caricature of the top-down process.

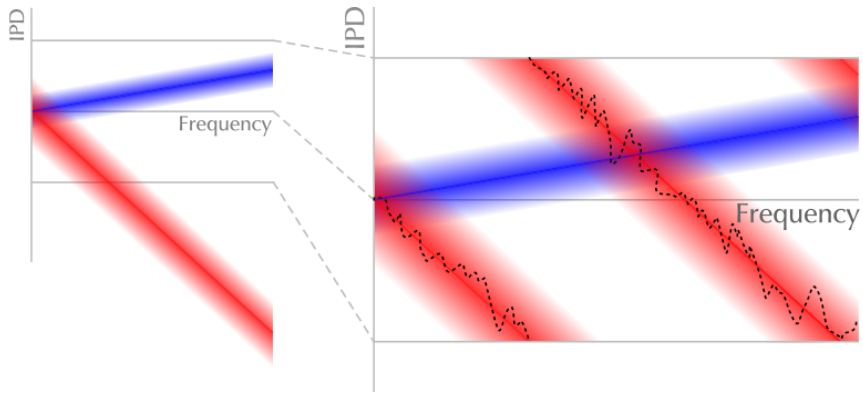


Figure 4.8: A caricature of our probabilistic model of interaural phase difference (IPD) as a function of frequency. On the left are the probability density functions (PDFs) of two competing model predictions, where the IPDs are centered around their respective $\omega\tau$ values. On the right, those PDFs are constrained to the interval $[-\pi, \pi)$ so that the likelihood of the observation (shown as a dotted line) can be evaluated under each.

Tobenkin (2009, Chap. 2) shows that the generalized cross-correlation and in particular the PHAT, can be easily transformed into an equivalent probabilistic cross-correlation using a von Mises noise model. Bradley and Kirlin (1984) and Knapp and Carter (1976) define a similar probabilistic cross-correlation, but under the much stricter assumptions of high SNR and signal and noise processes that are independent, Gaussian, and jointly wide-sense stationary. For completeness, we now derive the correspondence between equation (4.12) and equation (4.25) for multiple observations. Let $\bar{\Omega}$ be the region Ω but with all frequencies negated. Then through algebraic manipulation, we can transform the PHAT function to the log-likelihood of the von Mises distribution,

$$g(\tau) = \sum_{(\omega,t)\in\Omega\cup\bar{\Omega}} \frac{L(\omega,t)}{|L(\omega,t)|}\frac{R^*(\omega,t)}{|R^*(\omega,t)|} e^{j\omega\tau} \tag{4.39}$$

$$= \sum_{(\omega,t)\in\Omega\cup\bar{\Omega}} e^{j\angle L(\omega,t)} e^{-j\angle R(\omega,t)} e^{j\omega\tau} \tag{4.40}$$

$$= \sum_{(\omega,t)\in\Omega\cup\bar{\Omega}} \exp\left\{ j\angle \frac{L(\omega,t)}{R(\omega,t)} + j\omega\tau \right\} \tag{4.41}$$

$$= 2 \sum_{(\omega,t)\in\Omega} \cos(\phi(\omega,t) + j\omega\tau). \tag{4.42}$$

Equation (4.40) expresses PHAT as a relationship between just the phase portions of the two signals. Equation (4.41) combines the two signals' phases together in a single term. And equation (4.42) uses the definition of the IPD observation in equation (4.24) and groups each frequency with its matching negative frequency in the sum to convert the exponentials into cosines. The joint likelihood of a frame of IPD observations under the

von Mises model of equation (4.25) yields an expression of the same form as equation (4.42)

$$p(\phi(\Omega) \mid \tau) = \prod_{(\omega,t)\in\Omega} \mathrm{vM}_\theta(\phi(\omega,t); -j\omega\tau, \kappa) \tag{4.43}$$

$$\propto \prod_{(\omega,t)\in\Omega} \exp\left(\kappa \cos\left(\phi(\omega,t) + j\omega\tau\right)\right) \tag{4.44}$$

$$= \exp\Big(\kappa \sum_{(\omega,t)\in\Omega} \cos\big(\phi(\omega,t) + j\omega\tau\big)\Big) \tag{4.45}$$

$$= \exp\big(\tfrac{\kappa}{2} g(\tau)\big). \tag{4.46}$$

Because all of these transformations to $g(\tau)$ are monotonic, localization performed using PHAT in equation (4.8) is equivalent to choosing the maximum likelihood τ under the von Mises model. This result can also be extended to the case where κ varies over Ω. For example, Knapp and Carter (1976) discuss the Hannan-Thomson weighting (Hannan and Thomson, 1973) function that uses $\frac{1-\gamma^2}{\gamma^2}$ as the variance. This weighting is maximum likelihood under the restrictive assumptions that the authors make about the signal and noise.

4.3 Experiments

This section details two experiments that evaluate the performance of the probabilistic and region-based cross-correlation. The first experiment measures the localization ability of variants of the probabilistic algorithm in reverberation and additive noise. It also examines the effect of a mismatch between training and testing data. The second experiment measures the localization of the kernel-based normalized cross-correlation in a reverberant mixture of two sources.

Both experiments simulate directional sources by convolving a monaural utterance, recorded with a close-talking microphone to minimize noise and reverberation, with a binaural impulse response recorded separately in a real room. Specifically, these experiments use the classroom binaural room impulse responses (BRIRs) of Shinn-Cunningham et al. (2005), which were recorded using a KEMAR dummy head in a real classroom with a reverberation time of around 565 ms and a direct-to-reverberant energy ratio of around 10 dB. These experiments use the measurements taken in the middle of the classroom, with the source 1 meter from the listener at 0° and 75°. The recording equipment was taken down and re-setup on three different occasions, giving three different BRIRs at each angle. The maximum ITD for the dummy head is around 750 μs when the source is at 90°.

The anechoic impulse responses are actually the direct-path portion of these reverberant impulse responses. By setting all of the samples in the impulse responses to 0 after the first 10 ms, we are able to capture just the direct path and eliminate all of the early echoes and reverberation. To avoid artifacts, we used a 1 ms raised cosine window to smoothly fade out the signal between 9 and 10 ms. The noise added to these anechoic impulse responses was speech shaped, based on the average spectrum of the original recordings of all of the

utterances. The signal to noise ratio of the additive noise was 10.75 dB (relative to the average of both ears), to match the direct-to-reverberant ratio of the reverberant BRIRs.

4.3.1 Accuracy vs duration

The first experiment compared different variants of these maximum entropy distributions, including PHAT, for localizing single sources in simulated reverberation and additive noise. Twenty utterances were selected at random from the TIMIT dataset and each of these was spatialized at an azimuth of 0° and 75°. Each method was used to estimate the interaural time difference and the results were compared to the true ITD using the mean absolute error. The true ITD was estimated by passing 10 s of white noise through just the first 10 ms of the same impulse responses.

These localization methods give more accurate estimates when information is pooled over a longer observation, so their performance was compared as the amount of pooling was varied. Two different pooling methods were used, a point-based pooling related to PHAT-histogram and a probabilistic pooling as introduced in this chapter. In the point-based pooling, the maximum likelihood delay was selected at each time point and a histogram of such estimates was collected over the pooling interval. The final estimate was the delay with the largest histogram value. In probabilistic pooling, the likelihoods of all of the pooled frames were multiplied together and the maximum of that combined likelihood was selected.

The maximum entropy circular distributions that were compared use parameters estimated from different data. The first version used parameters estimated from a single utterance at 0° in the reverberant environment. The second used parameters estimated from the same utterance, at the same location, but using the anechoic additive noise condition described above. The third used parameters estimated from the same utterance, but located at 75° in reverberation. As was shown in section 3.1.3, the distribution of IPD residuals does differ with source azimuth, although the results of this experiment show that it has little effect on localization performance.

As shown in equation (4.46), the maximum likelihood delay under the von Mises distribution is equivalent to PHAT in this framework. In section 4.2.2, it was shown that the von Mises distribution is the first order maximum entropy circular distribution. Thus, in these experiments the first-order result is the same and as the performance of PHAT. In addition to the first-order result, the fifth-order result is shown.

4.3.1.1 Results

The results of these experiments can be seen in figure 4.9, which plots the mean absolute error of the localization methods as a function of the amount of pooling in the two conditions. Note that all of the methods perform very similarly to one another. The point-based estimate always performs significantly worse than the probabilistic estimate. Error decreases, monotonically as pooling increases, and the probabilistic method's error reaches below 20 μs for all methods when pooling over 300 ms.

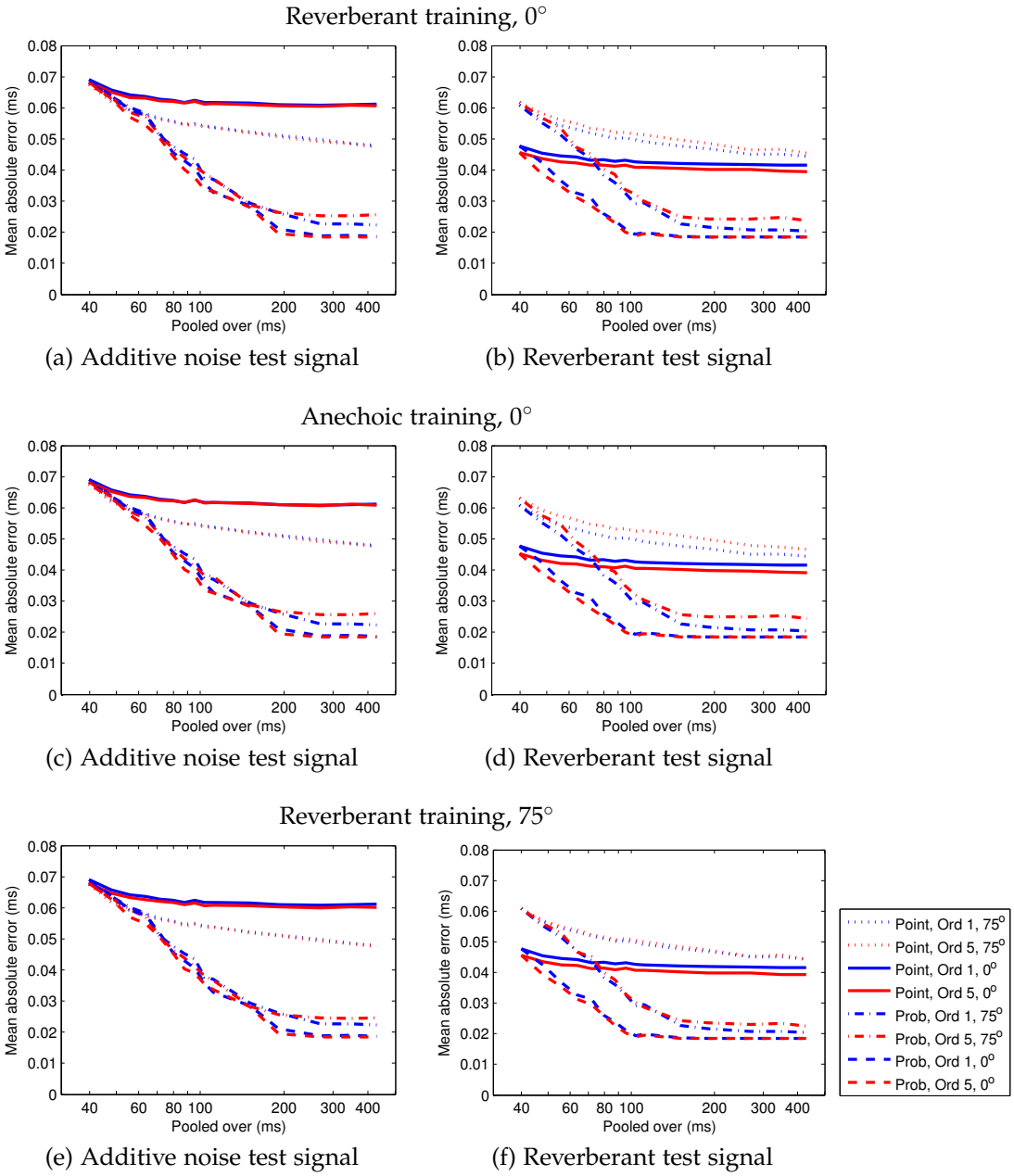


Figure 4.9: Mean absolute localization error using maximum entropy circular distributions evaluated in reverberant (right) and anechoic plus speech-shaped noise (left) conditions. Parameters were estimated from the indicated training signal types. Lines show performance for point and probabilistic estimates (color) for sources at 0° and 75° using first and second order circular distributions.

The training conditions of these algorithm does not change the results much at all. The plots in each row of figure 4.9 are very similar to one another. Also, the two shown model orders perform very similarly to one another, especially in the additive noise condition. In the reverberant noise condition, the higher order model helps the matched condition slightly, but hurts the unmatched condition. In general, however, the higher order model only seems to help in localizing the source at 0°.

With the probabilistic pooling, the source at 0° is more accurately localized than the source at 75° in reverberation, and they are localized equally well in additive noise. This is to be expected, as the IPD residual has lower variance for sources closer to 0°. Curiously, however, with point-based pooling, the performance of the two sources is reversed in the two conditions. This may be related to the advantage of the better ear compared to the average over both ears.

In reviewing the performance of these different methods, it became clear that almost all of the localization error came from a small number of intervals with very large errors, as opposed to many intervals with small errors. The segments with large errors were the segments containing little or no source energy. When pooling over short duration segments, this led to certain segments with a large error and others with very small error. As the pooling intervals were increased, there tended to be some target energy in every interval, leading to good localization in every interval, and decreasing the average error.

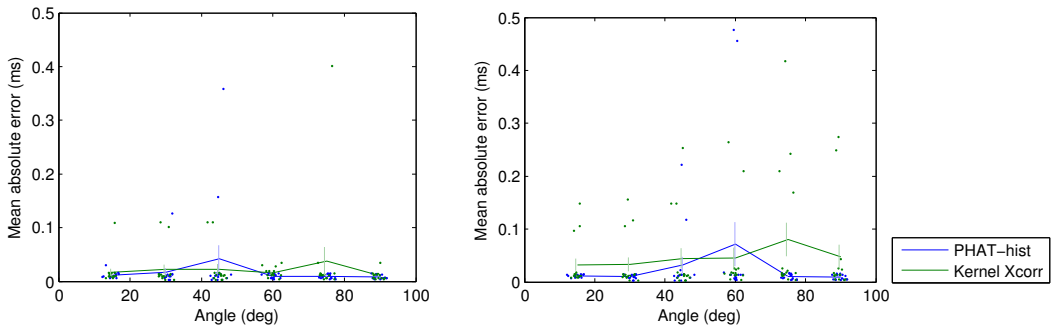
4.3.2 Localization in a mixture

To show the usefulness of the kernel-based normalized cross-correlation localization, we use it to localize multiple simultaneous sources. We compare it PHAT-histogram, which performs localization on each frame of a spectrogram separately, using PHAT, and selects the τ with the most weight as the source ITD at that frame. It then accumulates these point estimates into a histogram and selects the largest peaks as the sources' locations. Our region-based cross-correlation can be thought of as a generalization of the this technique. While PHAT-histogram uses "regions" of the spectrogram that are individual frames, our technique uses regions with frequency-dependent shapes. We combine our localizations in the same way, aggregating them in a histogram and selecting the largest peaks from that histogram.

In this task, two and three sources are mixed together in reverberation and the localization performance on all sources is measured as a function of separation angle. One source is always at 0°, and the other sources move from 15° to 90°. When there are three sources, the two off-axis sources are placed symmetrically at the same azimuth. Performance is again measured using the mean absolute error.

4.3.2.1 Results

The results of this experiment can be seen in figure 4.10. Each dot is the mean absolute error of the sources in a single mixture, the trend lines show the mean absolute error of all of the mixtures at each angle, and the error bars show 1 standard error. Again, most localizations are very close to the ground truth, but a few are very far off. When there are



(a) Two sources (b) Three sources

Figure 4.10: Mean absolute localization error for reverberant multiple-source mixtures. Error bars show 1 standard error.

three sources, the kernel cross-correlation tends to make more of these large mistakes than PHAT-histogram.

Note that the errors are relatively evenly distributed across angles. This is surprising because nearby sources have similar interaural parameters and one might expect that they would be confused more frequently. The increased incidence of confusions at small angles might be counteracted by the smaller errors caused by such confusions. See section 5.4.6 for further localization experiments involving these algorithms that show their limitations for sources separated by less than 10°.

When localization is successful, the performance of both of these systems is on par with that of humans. The median error in both systems' estimates is around 10–20 μs, which corresponds to an angle of approximately 2°. As discussed in section 2.3, human localization performance in anechoic environments is approximately 1–2° for sources directly ahead and as much as 10° for lateral sources.

4.4 Summary

This chapter discussed a number of approaches for localizing sound sources in azimuth, and presented two generalizations of previous localization methods. The first generalization expands the definition of cross-correlation from the time domain into the time-frequency domain, so that it can be calculated over arbitrarily shaped regions of a spectrogram. After showing that this localization is equivalent to a von Mises probability model, it introduces the second generalization, from a first order maximum entropy circular distribution to higher order distributions.

It also couched localization in a probabilistic framework, introducing a probability model for the interaural phase difference residual. The noise is zero mean, unimodal, and can be treated as independent for different time-frequency points. Thus the probability of the residuals in an arbitrary time-frequency region is simply the product of the probabilities

at all of the points in that region. By comparing the probabilities of different interaural time delays over an arbitrary region, the maximum likelihood delay can be chosen. This is essentially the "straightness" weighting of (Stern and Trahiotis, 1995), but arrived at from probabilistic principles.

There are a number of features of this model that recommend it over a traditional generalized cross-correlation approach. First, in localization experiments, the probabilistic approach estimated the true delay as well as PHAT in both reverberant and additive noise conditions. Second, the probabilistic approach lends itself to localizing multiple sources simultaneously because of its ability to aggregate information over arbitrarily shaped regions of the spectrogram. Third, because it is a true likelihood, it can be used in combination with other probabilistic methods for localization or separation. And fourth, it does not depend on any explicit assumptions of Gaussianity or wide-sense stationarity as previously proposed GCC methods do. Any implicit assumptions it makes are appropriate for use with speech in reverberation and additive noise.