A Probability Model for Interaural Phase Difference

Michael I. Mandel, Daniel P.W. Ellis

Department of Electrical Engineering
Columbia University, New York, New York
{mim,dpwe}@ee.columbia.edu

Abstract

In this paper, we derive a probability model for interaural phase differences at individual spectrogram points. Such a model can combine observations across arbitrary time and frequency regions in a structured way and does not make any assumptions about the characteristics of the sound sources. In experiments with speech from twenty speakers in simulated reverberant environments, this probabilistic method predicted the correct interaural delay of a signal more accurately than generalized cross-correlation methods.

1. Introduction

The human ability to localize sound sources surpasses that of even the best computer program, particularly in reverberant environments. Sound source localization could allow machine listeners to characterize and adapt to acoustic environments, and aid in filtering out unwanted sounds and more cleanly recognizing speech. Furthermore, humans are able to localize sound sources with just two sensors, while machines generally require large microphone arrays. In order to more accurately localize sound in reverberant environments, we propose a probability model for interaural phase differences at individual spectrogram points based on an empirical examination of measured binaural impulse responses.

This model considers time-frequency points in interaural spectrograms individually, facilitating the localization of multiple sources simultaneously. Even though each point contains only a small amount of information, combining the information at many points leads to accurate inferences of the location of sound sources. Furthermore, the interaural spectrogram, defined as the complex ratio between the spectrogram of the left and right ears, does not depend on any specific properties of the sound source such as Gaussianity or stationarity.

One problem that this approach addresses is the inherent ambiguity in phase differences. The difference in phase between the left and right ears at any frequency is only defined modulo 2π, so that a given phase difference could be caused by a number of delays between the two ears. A single delay, however, leads deterministically to a single phase difference. Previous methods have generally attempted to convert phase differences into delays, but we propose to model the probability of any particular delay given an observed phase difference, thus allowing the comparison of many different delays.

We test our probability model for interaural phase differences on the localization of single speakers in simulated reverberant and additive noise situations. For comparison, we also test the Phase Transform (PHAT), a generalized cross-correlation method, on the same simulations and show that our method outperforms PHAT on this task. We further show that our method is robust to a mismatch between the noise model and the observed noise distribution.

1.1. Previous work

There have been many previous efforts toward describing a perceptually meaningful binaural localization function. Knapp and Carter first introduced the generalized cross-correlation in [1]. Their method produces a point estimate of the time delay between two microphones by including a weighting function in a cross-correlation calculation. Under the assumption of stationary sources and uncorrelated noise at the two ears, they derive a maximum likelihood estimate of the delay. Another particularly useful weighting they introduce is the Phase Transform (PHAT), which whitens the two signals before cross-correlating them to provide a more sharply peaked correlation.

Approaching the problem from a more perceptual orientation, Stern and Trahiotis introduced a method for localizing sound sources in [2, 3]. They use cross-correlations between the two signals in many different frequency bands and then combine the estimates. The band limitation of the signals makes phase ambiguities quite evident, but all of the bands' correlations align at the true delay. The idea of coincident delays in the cross-correlation leads the authors to design a “straightness” measure for scoring alignments at different delays. Their straightness measure combined measurements in each frequency band by multiplying their correlations together, a procedure that can be justified when considering the problem in a probabilistic framework such as ours.

The classroom impulse responses we study were recorded by Shinn-Cunningham et al. in [4], in which the impulse responses were analyzed, along with the interaural time delay and interaural intensity difference. These impulse responses have been used for many experiments, including an investigation of human spatial audio perception [5] and neural modeling of binaural audition [6].

2. Probability model

For the purposes of deriving this model we will examine the situation where one sound source arrives at two spatially separate microphones or ears. Eventually this will generalize to the assumption that only a single source arrives at each time-frequency point in a spectrogram, but that different points could contain different sources.

Denote the sound source as $s(t)$, and the signals received at the left and right ears as $\ell(t)$ and $r(t)$, respectively. The two received signals will have some delay and some gain relative to the source, in addition to a disruption due to noise. For this model, we assume a convolutive noise process, because it fits our empirical observations, it is easy to analyze, and in general is it is very similar to the additive noise processes that other authors assume. The relation between the various signals is then,

$$\ell(t) = a_\ell s(t - \tau_\ell) * n_\ell(t) \tag{1}$$

$$r(t) = a_r s(t - \tau_r) * n_r(t). \tag{2}$$

Taking the ratio of the Fourier transform of both equations gives,

$$L(j\omega) = e^{a - j\omega\tau} R(j\omega) N(j\omega), \tag{3}$$

where $\tau = \tau_\ell - \tau_r$, $N(j\omega) = \frac{N_\ell(j\omega)}{N_r(j\omega)}$, and $a = \log \frac{a_\ell}{a_r}$. This equivalence assumes that τ is less than the length of the window over which the Fourier transform is taken, a condition easily met for dummy head recordings with moderately sized Fourier transform windows. For example, in our experiments the maximum delay was 0.75ms, and the window length was 32ms.

When localizing a signal, our goal is to infer a and τ from the observed signals $R(j\omega)$ and $L(j\omega)$, while avoiding distraction from the noise $N(j\omega)$. In order to accomplish this goal, however, it is necessary to study the behavior of the noise in both magnitude and phase for known a and τ. It is possible to separate the effects of phase and magnitude noise because for a given a and τ, the noise is described by

$$\log |N(j\omega)| = \log |L(j\omega)| - \log |R(j\omega)| - a \tag{4}$$

$$\angle N(j\omega) = \angle L(j\omega) - \angle R(j\omega) + \omega\tau + 2k\pi, \tag{5}$$

where k comes from the inherent 2π ambiguity in phases.

2.1. Data

In order to study this noise, we simulated speech in anechoic, reverberant, and additive noise situations by convolving anechoic speech samples with binaural impulse responses. We used speech from the TIMIT acoustic-phonetic continuous speech corpus [7], a dataset of utterances spoken by 630 native American English speakers. Of the 6300 utterances in the database, we chose 20 at random to use in our evaluation.

The anechoic binaural impulse responses came from Algazi et al. [8], a large effort to record head-related transfer functions for many different individuals. Impulse responses measurements were taken over the sphere surrounding subjects' heads at 25 different azimuths and 50 different elevations. The measurements we used were for the KEMAR dummy with small ears, although the dataset contains impulse responses for around 50 individuals.

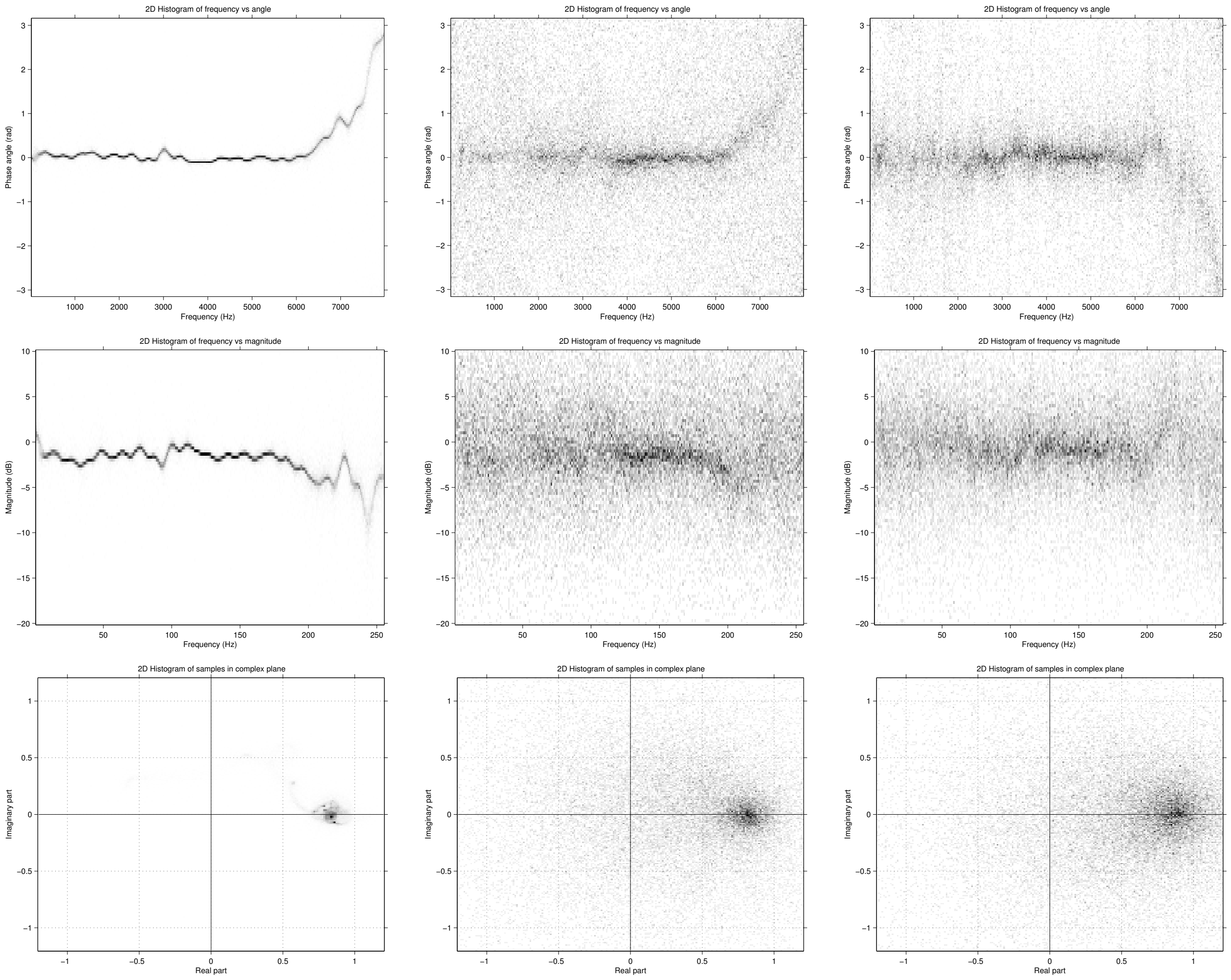
To compare with the anechoic impulse responses, we used impulse responses recorded in a real classroom made by Shinn-Cunningham et al. [4]. These measurements were also made on a KEMAR dummy, although a different actual dummy was used. Measurements were taken at four different positions in the classroom, three distances from the subject, seven directions, and three repetitions of each measurement. We used the measurements taken in the middle of the classroom at a distance of 1 meter from the subject.

2.2. Observations

See Figure 1 for some examples of our observations of the noise. These measurements were simulated from the same speech segment in three different conditions. First is the anechoic condition created by passing the speech through the anechoic impulse responses. Second is an additive noise condition, created by first passing the speech through the anechoic impulse responses and then adding independent, equal-power, speech shaped noise to each ear. The third condition is the reverberant environment, created by passing the speech through the classroom impulse responses.

The top plots show two-dimensional histograms of the phase noise at interaural spectrogram points as a function of their frequency. One can see that the noise is unimodal, with a mean and variance that are relatively consistent across frequencies. The highest frequencies show some deviation from this standard because of the well-studied notch in interaural magnitude caused by the pinna. We gloss over this detail and model the noise as identically distributed across frequencies.

The middle plots are similar to the top plots, but show magnitude noise in dB as a function of frequency. Again, the noise appears unimodal, with a consistent mean and variance across frequencies except for the notch at high frequencies.



Anechoic Additive noise Reverberation

Figure 1: Two-dimensional histograms of noise in anechoic, additive, and reverberant situations for 0° azimuth/elevation. Top: plotted as phase angle vs frequency. Middle: plotted as magnitude (in dB) vs frequency, Bottom: plotted in the complex plane. Note that columns one and two are derived from the same KEMAR, but the third column involves a different KEMAR with different pinna asymmetries.

The bottom plots show two-dimensional histograms of noise observations in the complex plane, i.e. the joint distribution of phase and magnitude. These histograms collapse the noise across all frequencies, and it can be seen that the distribution in the complex plane is unimodal, has much heavier tails than a Gaussian, and is located around the point $1 + 0j$, i.e. magnitude 1, angle 0.

2.3. Noise model

These observations indicate that to a first approximation, we can consider all interaural spectrogram points to be identically distributed. We make the further assumption that the errors at all interaural spectrogram points are independently distributed. The observations also indicate that the magnitude noise is well described by a lognormal distribution, and that the phase noise is well described by a circular probability distribution akin to the von Mises distribution.

For the rest of the paper, we will concern ourselves with the phase of the noise distribution, ignoring the magnitude. We now describe the model for the independent, identically distributed interaural phase measurements. For any single point in time and frequency, let

$$\Delta\phi \equiv \angle L(j\omega) - \angle R(j\omega) \quad (mod\, 2\pi). \tag{6}$$

Then $p(\Delta\phi \,|\, \tau)$ is peaked at $\omega\tau$ and periodic in 2π, leading to the approximation

$$p(\Delta\phi \,|\, \tau) = \exp\left(\sum_{k=0}^{K} a_k \cos^k(\Delta\phi - \omega\tau)\right). \tag{7}$$

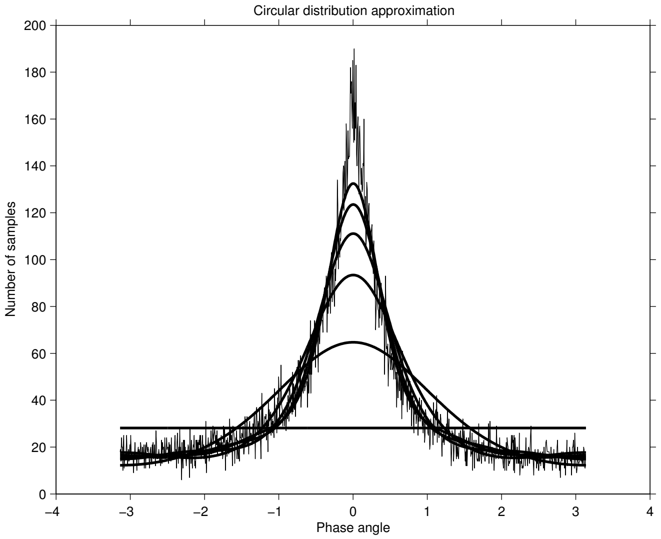


Figure 2: A histogram of measured noise angles in a reverberant environment with successive maximum entropy approximations superimposed.

By including more terms, the approximation can become as accurate as desired. This function is the maximum entropy distribution for a probability periodic in 2π with the first K moments specified.

It is also easy to match this model to observations by constructing a histogram of angles in $[0, 2\pi)$, taking the logarithm of the occupancies, and then taking the first K terms of the Fourier series. Since $\cos(K\theta)$ can be expressed as a Kth order polynomial in $\cos(\theta)$, the Fourier series representation is equivalent to the maximum entropy representation. See Figure 2 for an example of such a histogram and successive approximations to it.

Since the noise is independent and identically distributed for all points in the spectrogram, the joint probability of observations at multiple points is the product of their marginal probabilities.

$$p(\Delta\phi(\omega, \mathbf{t}) \,|\, \tau) = \prod_i p(\Delta\phi(\omega_i, t_i) \,|\, \tau). \qquad (8)$$

To combine the estimates of an entire spectrogram frame, then, one would need to multiply the probabilities of each point in that frame,

$$p(\Delta\phi(t) \,|\, \tau) = \prod_{n=0}^{N/2} p(\Delta\phi(2\pi jn/N, t) \,|\, \tau). \qquad (9)$$

Similarly, to combine the estimates of a frequency band over time, one could multiply across t, holding ω constant. But, the use of this probability model allows for the flexibility to combine probabilities across arbitrary regions of the spectrogram, including multiple frames and non axis-aligned regions.

In our experiments, we show that this probability model can be applied to the same problems previously attacked with the Generalized cross-correlation (GCC). In particular, we compare the performance of our model to that of the Phase Transform (PHAT) in terms of mean squared deviation from the true delay.

Using the same notation as above, the GCC assigns a score to each possible delay based on a weighted correlation. By taking the Fourier transform, this becomes

$$g(\tau) = \sum_{\omega} \psi(j\omega) L(j\omega) R^*(j\omega) e^{j\omega\tau}. \qquad (10)$$

The particular values of the $g(\tau)$ function do not matter, as only its maximum is used to choose the most likely delay, $\hat{\tau} = \arg\max_\tau g(\tau)$. One particular instance of the GCC is the PHAT, in which the weighting factor $\psi(j\omega)$ cancels the magnitudes of the left and right signals,

$$p(\tau) = \sum_{\omega} \frac{1}{|L(j\omega)||R(j\omega)|} L(j\omega) R^*(j\omega) e^{j\omega\tau}. \qquad (11)$$

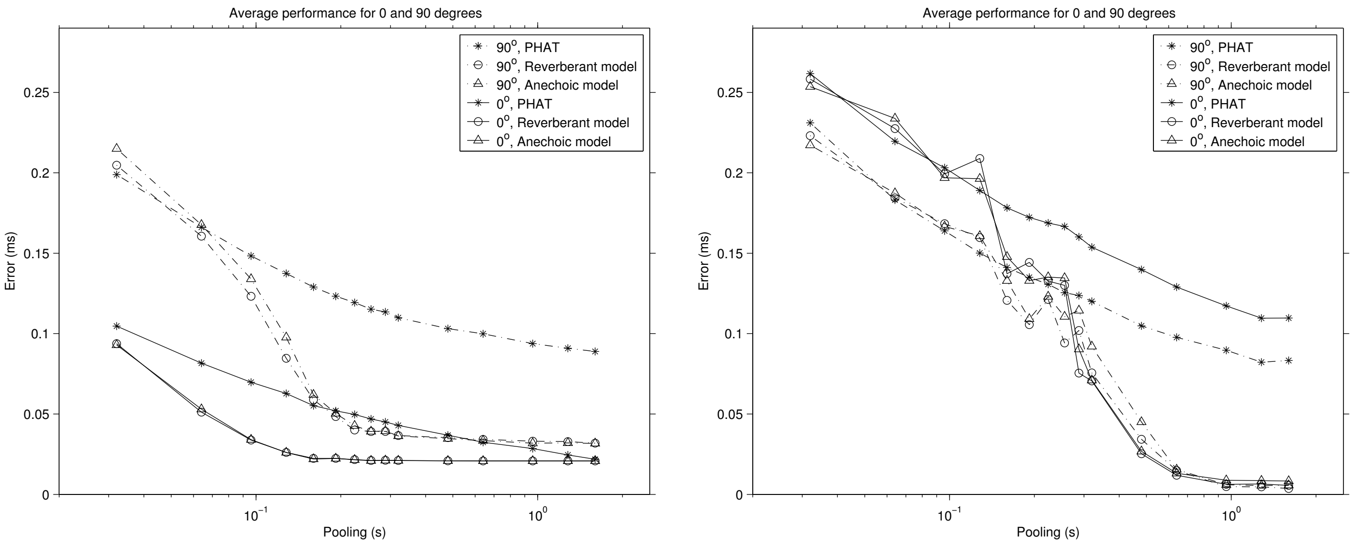
This whitening works well for broadband signals, but amplifies background noise when the target signal is in fact narrowband.

3. Experiments

To compare the performance of this probabilistic framework to GCC methods, we used them to localize single sources in simulated reverberation and additive noise. Twenty utterances from the TIMIT dataset were used and the errors averaged across utterances. Each method was used to estimate the interaural delay and the results were compared on the mean squared distance between the true values and these estimates.

The test in reverberation was performed using the classroom BRIRs from [4]. In particular, we used all of the recordings taken with the dummy head situated in the middle of the classroom with the sound source 1 meter from the head. This resulted in a total of 21 recordings from 7 different directions, from straight ahead to all the way to the right in increments of 15°.

The test in additive noise was performed using the anechoic HRIRs from [8]. In particular, we used the impulse responses at 0° elevation and between straight ahead and all the way to the right for a total of 13 recordings from 13 different directions. These directions were spaced more densely in front of the subject and more sparsely toward the side. The added noise was speech shaped, based on the average spectrum of the anechoic recordings of all of 20 of the utterances. The signal to noise ratio of the additive noise



Reverberation

Additive noise

Figure 3: Performance of the three localization methods in different noise situations. The three methods are PHAT, the probabilistic phase method using a distribution for reverberant noise, and the probabilistic phase method using a distribution for the anechoic situation.

was 10 dB (relative to the average of both ears), the measured direct-to-reverberant ratio in the classroom BRIRs.

The performance of three localization methods was measured in these two noise situations. All three methods give more accurate estimates when information is pooled over more observations, so the methods were compared when varying the amount of pooling. Since PHAT gives a single point estimate, pooling amounts to averaging the point estimates together. For the probabilistic estimates, pooling means multiplying the likelihoods. Two probabilistic methods were tested; for the first, the noise model was based on the reverberant recordings, while the second used a noise model based on the anechoic recordings.

The results of these experiments can be seen in Figure 3. The figure plots the mean squared error of the three localization methods as a function of the amount of pooling in the two noise situations. Performance on sounds straight ahead and fully to the right are plotted, intermediate angles have intermediate results and are omitted for clarity. As can be seen in the figure, all of the methods perform better with more pooling and, for short pooling intervals, they perform comparably to one another. With larger amounts of pooling, however, the probabilistic methods outperform PHAT. The behavior of the phase noise under additive noise conditions is similar enough to convolutive noise that the probabilistic methods still outperform PHAT. Curiously, localization at 90° is better than at 0° in additive noise, which may be related to the SNR advantage of the better ear compared to the average over both ears.

Matching the noise model to the actual noise properties gives the best performance, shown by the “reverberant model” traces in the left plot. However, even for a mismatched model (e.g. the “anechoic model” in the same plot), performance is degraded by only a small margin, and only for moderate pooling intervals. For mixtures of sources, it may become more important to use well-matched estimates of phase uncertainty to make accurate inferences of the azimuth associated with each spectrogram point.

4. Conclusions

We have introduced a probability model for the phase noise in interaural spectrograms. This noise model was developed through an examination of binaural recordings in anechoic and reverberant environments. The noise is zero mean, unimodal, and can be treated as independent for different spectrogram points. Thus the joint distribution of the noise in a region of a spectrogram is the product of the marginal noise distribution at each spectrogram point, meaning that posterior distributions of azimuth based on arbitrary sets of samples may be obtained by simply multiplying together the azimuth distributions derived from each point. This is essentially the “straightness” weighting of [3], 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 tests of simulated speech in reverberant environments, the probabilistic approach estimated the true delay with less mean squared error than PHAT. Second, even in additive noise, the situation for which GCC methods were designed, the probabilistic approach performs at least as well, if not better than PHAT. Third, the probabilistic approach makes no assumptions about the sources involved. And fourth, the probabilistic approach lends itself to localizing multiple source simultaneously because of its ability to aggregate information over arbitrarily shaped regions of the spectrogram.

In the future, we plan to integrate interaural magnitude differences into this localization system. We are also working on incorporating the entire model into an EM framework for localizing multiple sources. The basic idea of the EM approach to this problem is as follows. Since arbitrary time-frequency regions can be included in the calculation of probabilities under this model, one could group only the spectrogram points dominated by a single source into a calculation of the probability of the interaural delay of that source. Then with an estimate of the delays of the sources, one could assign each spectrogram point to each source in proportion to its probability of having originated from that delay.

5. Acknowledgments

The authors would like to thank Barbara Shinn-Cunningham and Richard Duda for sharing their binaural impulse responses with us. This work was supported by the Fu Foundation School of Engineering and Applied Science via a Presidential Fellowship and the National Science Foundation (NSF) under Grant No. IIS-05-35168. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

6. References

[1] C. H. Knapp and G. C. Carter, “The generalized correlation method for estimation of time delay,” in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 320–327, 1976.

[2] R. M. Stern, A. S. Zeiberg, and C. Trahiotis, “Lateralization of complex binaural stimuli: a weighted image model,” *Journal of the Acoustical Society of America*, vol. 84, pp. 156–165, 1988.

[3] R. M. Stern and C. Trahiotis, “Models of binaural interaction,” in *Handbook of Perception and Cognition, Volume 6: Hearing* (B. C. J. Moore, ed.), ch. 10, pp. 347–387, Academic Press, 1995.

[4] B. Shinn-Cunningham, N. Kopco, and T. J. Martin, “Localizing nearby sound sources in a classroom: Binaural room impulse responses,” *Journal of the Acoustical Society of America*, vol. 117, pp. 3100–3115, 2005.

[5] N. Kopco and B. G. Shinn-Cunningham, “Auditory localization in rooms: Acoustic analysis and behavior,” in *Proceedings of the 32nd International Acoustical Conference - EAA Symposium*, pp. 109–112, September 2002.

[6] B. G. Shinn-Cunningham and K. Kawakyu, “Neural representation of source direction in reverberant space,” in *Proceedings of the 2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79–82, October 2003.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT acoustic phonetic continuous speech corpus CDROM,” 1993.

[8] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, “The CIPIC HRTF database,” in *Proc 2001 IEEE Workshop on Applications of Signal Processing to Audio and Electroacoustics*, pp. 99–102, Oct 2001.