NEURAL REPRESENTATION OF SOURCE DIRECTION IN REVERBERANT SPACE

Barbara Shinn-Cunningham and Kosuke Kawakyu

Boston University Hearing Research Center
677 Beacon St.
Boston, MA 02215 USA
shinn@bu.edu, kosuke@mit.edu

ABSTRACT

Head-related impulse responses measured in a classroom are used to generate realistic "reverberant" inputs to a neural model of binaural processing. Results show that the instantaneous information in the midbrain provides relatively poor source direction information, even though human perception is only slightly affected by modest levels of reverberation. Methods for integrating information across time to improve localization accuracy give insight into why auditory processing is more robust to room effects than performance of most acoustic array processing schemes.

1. INTRODUCTION

Spatial hearing is important for monitoring the environment and enabling listeners to understand a signal in the presence of competing sources. Because of its behavioral and theoretical importance, there has been intense interest in spatial hearing for over a century. However, most past studies of spatial hearing ignored the effects of realistic room reverberation on spatial acoustic cues. Algorithms for combining information across multiple receivers have been designed to perform tasks like those executed by the spatial hearing system. However, whereas spatial hearing abilities are only modestly degraded with room reverberation [1], many multi-microphone algorithms degrade drastically [2].

This study examines how classroom echoes influence the representation of interaural time differences (ITD; a dominant acoustic cue for source direction) in a model population of brainstem neurons and considers methods for integrating noisy, instantaneous source direction estimates. This work may give insight into why neural processing is relatively insensitive to reverberation, ultimately leading to more robust microphone array processing algorithms that are better able to cope with room reverberation.

2. MODELED BRAINSTEM RESPONSES

2.1. Simulating the Signals Reaching the Ears

Head-related impulse responses (HRIRs; see [3]) were measured in a classroom (rough dimensions 5 m x 9 m x 3.5 m) to simulate reverberant signals. Hearing-aid microphones inserted into the blocked ear canals of a Knowles Electronics manikin (KEMAR) recorded acoustic responses. The resulting reverberant impulse responses (length 740 ms; measured using a Maximum-Length Sequence) include the normal echoes and reverberation in the room. HRIRs were measured with KEMAR located in the center of the room and in the corner (with his back and left side within about 20 cm of a wall). Measurements were taken with the acoustic source at a distance of one meter and various azimuth angles from 0° to 90° to the right relative to KEMAR's head (in the horizontal plane containing the ears). The "center" HRIRs were time-windowed to remove all echoes and reverberation, creating "pseudo-anechoic" HRIRs used to simulate sources in anechoic space. To simulate the appropriate left- and right-ear signals reaching the listener's ears, the desired sound source (a one-second long sample of Gaussian noise) was convolved with the appropriate left- and right-ear HRIR.

2.2. Simulating Neural Responses

To a first-order approximation, each auditory nerve fiber (ANF) in the human auditory system responds to a band-passed version of the acoustic input signal, with the center frequency of the bandpass filter varying from fiber to fiber. To model neural processing of ITD, left and right ear signals simulated using HRIRs were processed through a realistic ANF model [4]. This model mimics some of the nonlinear effects observed in real ANFs, including adaptation to emphasize onset responses. The model output is the instantaneous, time-varying probability of observing a neural spike in a fiber of a specified center frequency given a particular input signal.

Highly specialized neural circuitry in the brainstem (specifically, in the medial superior olive or MSO) is thought to be the initial site of significant ITD processing in the mammalian auditory pathway [5]. MSO neurons are "tuned" (respond preferentially) to both input frequency and ITD because they act as interaural "coincidence detectors," generating output spikes if they receive nearly-simultaneous neural spikes from frequency-matched ipsilateral and contralateral inputs. The delay from the time an acoustic signal reaches an ear to the time a corresponding spike reaches an MSO neuron differs for the ipsilateral and contralateral inputs; as a result, different MSO neurons have different "best" (preferred) ITD values, equal to the ITD that compensates for any differences in ipsilateral and contralateral neural transmission delays to the MSO. Many binaural models approximate the output of the MSO coincidence-detector cells by computing the cross-correlation of frequency-matched left and right ANF inputs. The magnitude of this function at a particular cross-correlation delay predicts the expected firing rate of a neuron with a given best ITD; however, because it does not include

any refractory effects, it overestimates the firing rate at the onset, where the ANF inputs are very high.

A running cross-correlation of frequency-matched ANF responses was computed within a rectangular time window (length exactly four cycles of the center frequency or CF). Overlap from sample to sample was 50%. The resulting time-varying vector represents the instantaneous population response of MSO neurons tuned to different ITDs. The value of each vector element estimates the instantaneous firing rate of the MSO neuron tuned to a particular ITD.

2.3. MSO Simulation Results

Figure 1 shows the time-varying outputs of a population of MSO neurons receiving inputs from ANFs tuned to 547 Hz in response to simulated noise sources. The abscissa shows post-stimulus presentation time. The ordinate corresponds to the best ITD of each neuron in the 547-Hz population. Within each panel, the firing rate of one model neuron (as a function of time) is given by the image intensity in the corresponding horizontal line (with dark representing no firing and light representing high firing rates). Columns show results for anechoic, center, and corner acoustic environments (left, center, and right columns, respectively), while rows show results for a source at azimuth angles of 0˚, 45˚, and 90˚ relative to the median plane (top, center, and bottom rows, respectively).

For a source in anechoic space (left column of Figure 1), the instantaneous neural responses change little with time. The neurons whose best ITD corresponds closely to the expected ITD (given the simulated azimuth angle in the different rows) show the greatest activity and only neurons whose best ITD is within roughly 100 μs of the expected peak show significant firing. The only exception to this occurs for the 90-degree source (bottom row), where a second peak of activity is just visible for an ITD of –1 ms (left ear leading). This second activity peak is due to an inherent phase ambiguity caused by the peripheral band-pass

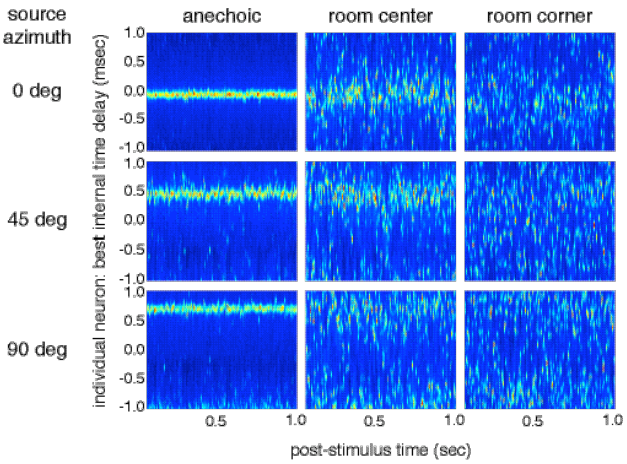

Figure 1. *Instantaneous output of a population of interaural-time-delay sensitive neurons tuned to 547 Hz. Simulated, one-second-long Gaussian-noise sources were located one meter from the center of the head at various azimuths (rows) in different acoustic environments (columns).*

filtering of the ANF inputs. Due to the relatively narrow bandwidth of the modeled ANF inputs, an external ITD of +800 μs (right ear leading) causes significant correlation in the narrowband inputs at multiples of (1/547 Hz) or 1.83 ms away from the "true" peak, producing a secondary peak at approximately -1 ms (left ear leading).

A source simulated at the center of the room (center column) generates greater variation in the peak MSO activity over time. Additionally, at each time instant, the peak activity tends to be smaller in magnitude and a larger number of neurons fire. In essence, the echoes and reverberation lead to a broader, more distributed neural response in which the mean activity centers around neurons tuned to the "correct" ITD, but the activation is much more variable. Finally, the secondary peak response near -1 ms (left ear leading) for a 90-degree source (bottom row) is much more evident in reverberant than in anechoic results. Population responses for the corner position (right column) are similar to those for the center position, but show even larger fluctuations, smaller peak activity, and a more diffuse population response.

The trends in Figure 1 also arise for other CFs: anechoic responses are consistent over time with narrow activity peaks and reverberant responses show greater variability over time with less defined, smaller peak activity. However, as CF increases, the phase locking of the ANF inputs decreases, resulting in population activity that changes less with source azimuth and secondary peaks that are more likely to fall within the physiologically-plausible ITD range.

3. INSTANTANEOUS ESTIMATES OF LOCATION

Many psychophysical studies suggest that for narrowband signals, subjects are sensitive to stimulus IPD rather than stimulus ITD. Only by combining IPD estimates across frequency is the "true" source ITD perceived [6]. This observation suggests that each population of MSO neurons (tuned to a particular CF) should yield an estimate of source IPD and motivates the approach detailed here in which, at each time instant, the responses of all MSO neurons with the same frequency tuning are combined to form an IPD estimate.

Previous analysis of a similar model MSO population found the maximum likelihood estimate (MLE) of source angle for a narrowband input with center frequency f [7]. When f is relatively low and best ITDs are equally distributed from –T to +T ms, the MLE equals the phase angle of a weighted sum of complex values, with each complex value representing the response of one neuron:

$$\hat{\Delta}_i(f) = \angle \sum_{\Delta_m=-T}^{T} L_{i,\Delta_m}(f) e^{j2\pi f \Delta_m}, \qquad (1)$$

where $L_{i,\Delta_m}(f)$ is the instantaneous firing rate at time i for the neuron whose best ITD is Δ_m.

Eq. (1) treats each neuron's response as a complex value whose magnitude is the instantaneous firing rate and whose phase is the best IPD of the neuron. While the phase of the complex-valued sum in Eq. (1) estimates source IPD around frequency f at time i, the magnitude of the sum also contains information. Specifically, if at a particular time instant the neural response is confined to neurons whose best IPD is

near □, then the resultant value will point to an IPD of □ and have a large magnitude. However, if the activity is spread over many neurons with different best IPDs, the summed value will have a relatively small magnitude, corresponding to a relatively low *a posteriori* probability of the estimated IPD leading to the observed population response at that time instant, even though that IPD is the MLE.

The current MSO model results are very similar to the form of MSO responses assumed in the derivation of Eq. (1); therefore, Eq. (1) should give a near-optimal estimate of IPD for the current results. Figure 2 plots the complex-valued resultant vectors (one per time sample) for an MSO population tuned to 547 Hz in response to a one-second-long Gaussian noise. Time samples are spaced every 2 periods of the CF (2/547 Hz or 3.6 ms in Figure 2). The nine panels are organized as in Figure 1, with columns representing acoustic environment (anechoic, center, corner) and rows representing source azimuth (0˚, 45˚, 90˚). Because the ANF model emphasizes stimulus onset and the cross-correlation approximation does not include any refractory period, the instantaneous MSO output rates are extremely large at onset (larger than the true physiological onset responses). In order to show both onset (black exes) and ongoing responses (gray dots) on the same plot, ongoing response magnitudes (gray dots) were scaled. In anechoic space, the ongoing responses were scaled up by a factor of five; in the center and corner conditions, the ongoing samples were scaled up by a factor of 13.

In general, onset IPD estimates (black exes) are similar in all three acoustic environments and vary systematically with source azimuth. In the anechoic environment (left column), ongoing IPD estimates are consistent from sample to sample

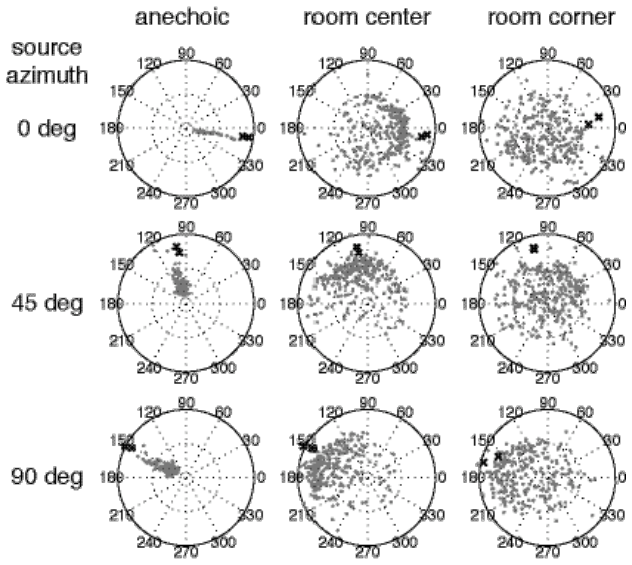

Figure 2. *Instantaneous estimates of source IPD from 547-Hz MSO neural responses shown in Figure 1 (panels organized as in Figure 1). Each point represents one estimate, with polar angle showing estimated IPD (radial length is a measure similar to vector strength). Initial three estimates are shown as black exes (post-stimulus onset times of 0, 3.6, and 7.2 ms). Subsequent estimates (every 3.6 ms for a one-second duration) are gray circles.*

(points fall roughly along a radial line in the same direction as the onset responses). Reverberation causes temporal fluctuation in the ongoing phase angle estimates, producing points that are spread randomly around the onset IPD for a source in a particular direction. This spread is greater for the corner than the center environment. In addition to a larger spread, estimates in the corner condition show a small displacement in IPD relative to the corresponding anechoic cases. For instance, even the onset IPD values are displaced slightly relative to the onset IPD estimates in the anechoic and center conditions. The main acoustic feature differentiating the corner and center conditions is the presence of early, intense reflections in the corner condition, which may cause these systematic shifts in estimated IPD.

4. INTEGRATION OF INFORMATION ACROSS TIME

There is ample evidence for "sluggishness" in the binaural processing system; listeners integrate IPD information over time and have difficulty perceiving rapid IPD fluctuations. Results in Figure 2 show that a source at a fixed location in a reverberant environment produces instantaneous estimates of IPD that are not very reliable; in fact, sluggishness may reflect the need to integrate noisy, instantaneous estimates to achieve acceptable localization accuracy in most typical listening environments. Two schemes for integrating IPD estimates over time are considered in this section.

A straightforward method for combining instantaneous estimates of direction computes the mean IPD across time samples. If the set of IPD estimates {□ᵢ} are independent, identically-distributed random variables with an expected value equal to the "true" source IPD, then the optimal unbiased, linear estimator (in the least-mean-square error sense) is given by the average of the instantaneous estimates. The mean phase value, averaged over the one-second-long signal duration, is shown in the top row of Figure 3 (error bars show the standard deviation in {□ᵢ}). Results for CFs of 345 Hz, 547 Hz, and 1094 Hz are shown in the left, center, and right columns for sources from 0˚ to 90˚. To simplify across-frequency comparisons, results were converted back to ITD (unwrapping IPD estimates). Exes, circles, and squares show results for anechoic, center, and corner conditions, respectively. The middle row plots the standard deviation in {□ᵢ}. The bottom row shows results of an alternative method, discussed below.

Results show that the across-time mean IPD is similar for anechoic and center conditions (within one standard deviation). However, there are small departures between estimates in the corner and anechoic conditions, probably due to the early, intense echo from the nearby walls in the corner condition, which are qualitatively different from the diffuse, random-direction reflections that dominate in the center room condition. The dependence of IPD estimates on source azimuth is similar across frequency.

The second row in Figure 3 shows that variability in {□ᵢ} is an order of magnitude larger in the presence of room reverberation than in anechoic space. This finding suggests that the amount of time averaging required to yield stable estimates of IPD depends on the level of the direct sound relative to the amount of reverberation (and thus on both listening environment and on source distance relative to the listener). Variability in {□ᵢ} varies with CF due to tradeoffs

between phase-locking (which increases variability with CF) and the maximum IPD error that can be observed (which limits variability at higher CFs; e.g., an IPD difference of π radians translates to an ITD discrepancy of approximately 500 μs for a CF of 1094, but nearly one ms for a CF of 547 Hz). Thus, the standard deviation is smaller for 1094 Hz than for 547 Hz, even though phase locking is worse.

If the values of $\{\theta_k\}$ are independent and identically distributed, their mean value is the optimal linear estimator; however, this approach ignores the reliability of each instantaneous IPD estimate. To weight the phase in each estimate proportional to the reliability, one can simply sum the instantaneous complex values in Eq. (1) and form

$$\hat{\Delta}(f) = \angle \sum_i \sum_{\Delta_m=-T}^{T} L_{i,\Delta_m}(f) e^{j2\pi f \Delta_m}, \qquad (2)$$

which automatically emphasizes information at times when instantaneous MSO activity if dominated by neurons with similar best IPD. The third row in Figure 3 plots the phase of the complex-valued sum of instantaneous estimates (summed over the one-second-long noise sample). These results are similar to the simpler mean IPD results; however, the difference between the anechoic and reverberant estimates is slightly smaller in some cases. In other words, taking into account how diffuse the MSO response is at a given time instant yields small improvements in estimation accuracy. However, significant estimation errors remain whenever there are early, intense echoes in a room.

5. SUMMARY AND CONCLUSIONS

In anechoic space, the response of model ITD-sensitive MSO neurons is remarkably stable over time and each individual time instant provides a reliable estimate of source direction. However, in a room, MSO activity fluctuates from instant to instant and is spread over a larger population of neurons. Instantaneous estimates of source direction based on MLE estimates of source angle are nearly the same in anechoic and reverberant space. However, temporal fluctuations in these estimates are an order of magnitude larger in the presence of normal room reverberation compared to in anechoic space. Simple methods for integrating information over time yield reasonably accurate estimates of source direction by averaging out the random variations from sample to sample. Small improvements in estimation accuracy can be realized by considering the reliability of the IPD estimate at each time instant in addition to its value. Although the estimation techniques presented here do not incorporate these effects, the human system also is thought to include nonlinear inhibition when processing interaural differences that further emphasizes the weight given to instantaneous direction estimates that are reliable, such as at onset [8].

In a room, even noise signals with small temporal intensity fluctuations yield IPD values that never settle into a "steady state" response, even though a signal such as a sinusoid would produce IPD estimates that rapidly settle to a steady-state value. The fact that most acoustic array processing techniques treat each time sample identically without considering how the reliability of directional cues varies over time may be a fundamental reason why the neural system is more robust in the face of room reverberation.

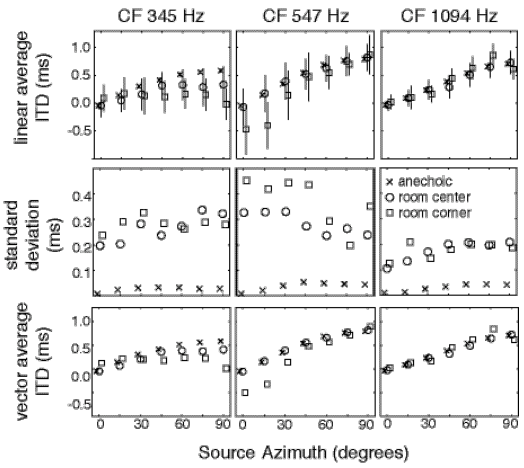

Figure 3. *ITD estimates integrated over time. Columns show results for different frequencies. Top row shows mean ITD estimates. Middle row plots the across-time standard deviation in the estimates. Bottom row shows ITD estimated from phase angle of complex-value sum estimation method.*

6. ACKNOWLEDGEMENTS

Portions of this work were supported by the Air Force Office of Scientific Research and the National Institutes of Health.

7. REFERENCES

[1] B. G. Shinn-Cunningham, "Learning reverberation: Implications for spatial auditory displays," Proceedings of the International Conference on Auditory Displays, pp. 126-134.

[2] J. E. Greenberg, J. G. Desloge, and P. M. Zurek, "Evaluation of array-processing algorithms for a headband hearing aid," *Journal of the Acoustical Society of America*, vol. 113, pp. 1646-1657, 2003.

[3] S. Carlile, *Virtual Auditory Space: Generation and Applications*. New York: RG Landes, 1996.

[4] M. G. Heinz, X. Zhang, I. C. Bruce, and L. H. Carney, "Auditory nerve model for predicting performance limits of normal and impaired listeners," *Acoustic Research Letters Online*, vol. 2, pp. 91-96, 2001.

[5] P. X. Joris, P. H. Smith, and T. C. T. Yin, "Coincidence detection in the auditory system: 50 years after Jeffress," *Neuron*, vol. 21, pp. 1235-1238, 1998.

[6] C. Trahiotis and R. M. Stern, "Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase," *Journal of the Acoustical Society of America*, vol. 86, pp. 1285-1293, 1989.

[7] H. S. Colburn and S. K. Isabelle, "Models of binaural processing based on neural patterns in the medial superior olive," in *Auditory Physiology and Perception*, Y. Cazals, K. Horner, and L. Demany, Eds. Oxford: Pergamon Press, 1992, pp. 539-545.

[8] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *Journal of the Acoustical Society of America*, vol. 106, pp. 1633-1654, 1999.