

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 physiologically-based model of binaural processing in the brainstem. Results show that instantaneous information provides relatively poor source direction information, even though human perception is only slightly affected by reverberation. However, analyses of the distortion of spatial cues caused by real room reverberation demonstrate that appropriate combination of noisy cues over time can increase the reliability of source direction estimates.

1. INTRODUCTION

Spatial hearing is important for monitoring the environment and enabling listeners to understand a signal in the presence of competing sources. However, most past spatial hearing studies have ignored how room reverberation affects the spatial acoustic cues available at the ears. This question has practical significance in that spatial hearing abilities are only modestly degraded with room reverberation [1], but 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. While based on past approaches, this work differs in that it evaluates the reliability of spatial information extracted by neural processing of stimuli with realistic reverberation. Results may give insight into what spatial information is robust in everyday environments that contain echoes and thus how the auditory system achieves relatively accurate spatial performance despite the effects of room acoustics.

2. MODELED BRAINSTEM RESPONSES

2.1. Simulating the Signals Reaching the Ears

Reverberant impulse responses were measured in a classroom (rough dimensions 5 m x 9 m x 3.5 m) by inserting hearing-aid microphones into the blocked ear canals of a Knowles Electronics manikin (KEMAR). Measured reverberant impulse responses (length 740 ms; measured using a Maximum-Length Sequence), which will be referred to simply as head-related impulse responses (HRIRs), included 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 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, producing HRIRs used to create “pseudo-anechoic” stimuli. In all conditions, HRIRs were convolved with the desired sound source (a one-second long sample of Gaussian noise) to generate left- and right-ear signals.

2.2. Simulating Neural Responses

To a first-order approximation, each auditory nerve fiber (ANF) in the auditory system responds to a band-passed version of the acoustic input, with the center frequency of the bandpass filter varying from fiber to fiber. To model neural processing of ITD, HRIR-processed left and right ear signals were presented to a realistic ANF model [3]. This model mimics some 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.

Specialized neural circuitry in the brainstem (i.e., in the medial superior olive or MSO) is thought to be the initial site of significant ITD processing in the mammalian auditory pathway [4]. MSO neurons are “tuned” (respond preferentially) both to a particular input frequency (the characteristic frequency or CF) and a particular 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 responses are very high.

A running cross-correlation of CF-matched ANF inputs was computed over a rectangular time window (length equal

to four cycles of the CF, with a sample-to-sample window overlap of 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

Fig. 1 shows the time-varying outputs of a population of MSO neurons (CF of 547 Hz) in response to noise sources from various directions presented in different environments. The abscissa shows post-stimulus time. The ordinate shows the best ITD of each neuron in the 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 top, center, and bottom rows show results for a source at azimuth angles of 0°, 45°, and 90°, respectively.

For a source in anechoic space (left column of Fig. 1), the instantaneous neural responses change little with time. The neurons whose best ITD corresponds to the expected ITD (given the simulated azimuth angle) show the greatest activity and only neurons whose best ITD is within roughly 100 μ s 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 filtering of the ANF inputs. Due to the relatively narrow bandwidth of the modeled ANF inputs, an external ITD of +700 μ s (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 -1.1 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 and 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, primarily due to the interaural decorrelation caused by the reverberation. 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 Fig. 1 also arise for other CFs: anechoic responses are consistent over time with narrow activity peaks; 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 anechoic signals, subjects are sensitive to interaural phase differences (IPDs), not ITD. Only by combining IPD cues across frequency is the “true” source ITD perceived [5]. 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 the response of a similar MSO model population to an anechoic, narrowband source with center frequency f describes how to calculate the maximum likelihood estimate (MLE) of the narrowband source IPD [6]. One can show that under certain assumptions (e.g., f is relatively low; best ITDs are equally distributed over a symmetrical range from $-T$ to $+T$ ms; total activity is constant with source IPD; etc.), the MLE for source IPD can be written as a weighted sum of complex values, with each complex value representing the response of one neuron:

$$\hat{\Gamma}_i(f) = \sum_{\varphi_m = -T}^T L_{i,\varphi_m}(f) e^{j2\pi\varphi_m}, \quad (1)$$

where $L_{i,\varphi_m}(f)$ is the instantaneous firing rate at time i for the neuron whose best ITD is φ_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 φ , then the resultant value will point to an IPD of φ 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

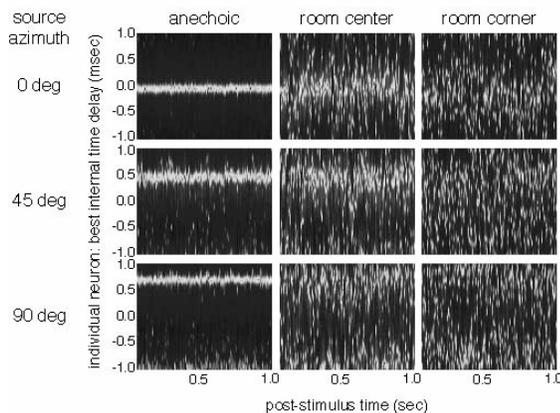


Figure 1. Instantaneous output of a population of interaural-time-delay sensitive neurons tuned to 547 Hz. Simulated, one-second-long Gaussian-noise sources were located one meter from the center of the head at various azimuths (rows) in different acoustic environments (columns).

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, nonetheless, the MLE.

The current MSO model responses to broadband noise are very similar to the assumptions leading to Eq. (1); therefore, Eq. (1) should give a near-optimal estimate of the IPD around frequency f for broadband sources in reverberant space. Fig. 2 plots the complex-valued resultant vectors (one per time sample) for a 547-Hz MSO population 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 Fig. 2). The nine panels are organized as in Fig. 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 up by a factor of five in anechoic space and by a factor of 13 in the reverberant conditions.

In general, onset IPD estimates (black exes) are similar in all environments and vary systematically with source azimuth. In the anechoic environment (left column), ongoing IPD estimates are consistent from sample to sample (all points, including onset responses, fall roughly along a radial line). 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, in the corner condition the mean estimated IPD differs from the anechoic cases. For instance, even the onset IPD values are

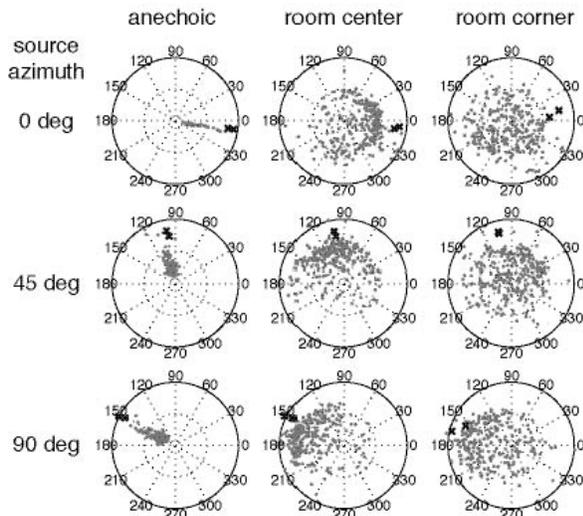


Figure 2. Instantaneous estimates of source IPD from 547-Hz MSO neural responses (panels as in Fig. 1). Each point is one estimate, with polar angle showing estimated IPD (radial length is a measure similar to vector strength). Initial three estimates are shown as black exes (post-stimulus onset times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

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

Listeners have difficulty perceiving rapid ITD fluctuations, suggesting that the auditory system integrates spatial information over time. Results in Fig. 2 show that a source at a fixed location in a reverberant environment produces relatively unreliable instantaneous estimates of IPD, reflecting a need to integrate noisy, instantaneous estimates to achieve acceptable localization accuracy. Two schemes for integrating IPD estimates over time are considered here.

A straightforward method for combining instantaneous estimates of direction computes the mean IPD across time samples. If the set of IPD estimates $\{\hat{\theta}_i\}$ are independent, identically-distributed random variables with an expected value equal to the “true” source IPD, the optimal unbiased, linear estimator (in the least-mean-square error sense) equals 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 Fig. 3 (error bars show the standard deviation in $\{\hat{\theta}_i\}$). Results for CFs of 345 Hz, 547 Hz, and 1094 Hz are shown in the left, center, and right columns. 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 $\{\hat{\theta}_i\}$. 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. 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 Fig. 3 shows that variability in $\{\hat{\theta}_i\}$ is an order of magnitude larger in a reverberant room compared to in anechoic space. This suggests that the amount of temporal 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). Variability in $\{\hat{\theta}_i\}$ depends on CF because of 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 causes an ITD discrepancy of roughly 500 μ s 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 less precisely at the higher frequency.

If the values of $\{\hat{\theta}_i\}$ 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

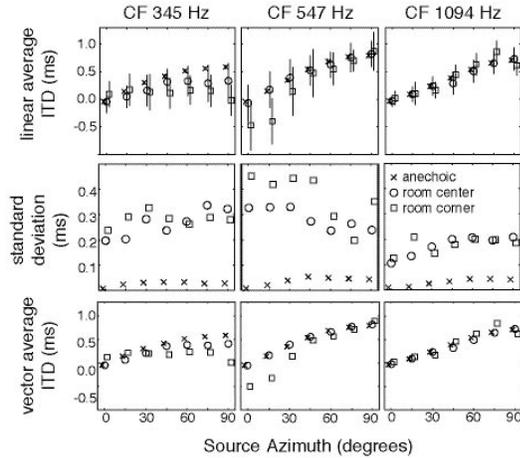


Figure 3. *ITD estimates integrated over time.* Columns show results for different frequencies. Top row shows mean ITD estimates. Middle row plots the across-time standard deviation in the estimates. Bottom row shows ITD estimated from phase angle of complex-value sum estimation method (Eqtn. 2).

$$\hat{\Lambda}(f) = \prod_i \prod_{\square_m = \square^T}^T L_{i, \square_m}(f) e^{j2\pi \square_m}, \quad (2)$$

which automatically emphasizes information at times when instantaneous MSO activity is dominated by neurons with similar best IPD. The third row in Fig. 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

Nearly all past studies of how the auditory system estimates source direction have ignored the effects of room acoustics on the cues reaching the ears. These results show that the effect of reverberation on the representation of spatial cues in the brainstem is pronounced. 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 calculated as the MLE of the source IPD have nearly the same expected value for anechoic and reverberant inputs; however, temporal fluctuations in these estimates are an order of magnitude larger in a room than 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. Of course, the auditory system is known to have spatial inhibition that emphasizes spatial cues in a stimulus onset [7]. However, recent results show that inhibition of later-arriving spatial cues occurs even in the absence of a strong onset [8, 9], as if the spatial auditory pathway employs a more general inhibitory mechanism in which strong, reliable spatial cues lead to delayed inhibition, particularly of spatial cues inconsistent with past information. While even simple methods that use knowledge of the estimated reliability of each instantaneous piece of spatial information yield small improvements in estimated direction, the auditory system appears to use a relatively sophisticated, nonlinear processing scheme that takes into account the reliability and consistency of spatial information over time. 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 relatively more robust than many machine algorithms in the face of room reverberation.

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," *Proc of Int Conf Aud Disp*, Atlanta, GA, pp. 126-134, 2000.
- [2] J. E. Greenberg, J. G. Desloge, and P. M. Zurek, "Evaluation of array-processing algorithms for a headband hearing aid," *J Acoust Soc Am*, vol. 113, pp. 1646-1657, 2003.
- [3] 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.
- [4] 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.
- [5] C. Trahiotis and R. M. Stern, "Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase," *J Acoust Soc Am*, vol. 86, pp. 1285-1293, 1989.
- [6] 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.
- [7] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The precedence effect," *J Acoust Soc Am*, vol. 106, pp. 1633-1654, 1999.
- [8] R. M. Dizon, "Behavioral and Computational Modeling Studies of the Precedence Effect in Humans," in *Biomedical Engineering*. Boston: Boston University, 2001.
- [9] S. Santarelli, N. Kopco, and B. G. Shinn-Cunningham, "Distance judgements of nearby sources in a reverberant room: Effects of stimulus envelope," *J Acoust Soc Am*, vol. 107, 2000.