A Biologically-Inspired Monaural Sound Localizer

John G. Harris, Chiang-Jung Pu, and Jose C. Principe
Computational Neuro-Engineering Laboratory
Department of Computer and Electrical Engineering
University of Florida
Gainesville, FL 32611

Abstract

We describe the first single microphone sound localization system and its inspiration from theories of human monaural sound localization. Reflections and diffractions caused by the external ear (pinna) allow humans to estimate sound source elevations using only one ear. Our single microphone localization model relies on a specially shaped reflecting structure that serves the role of the pinna. Specially designed analog VLSI circuitry uses echo-time processing to localize the sound. A CMOS integrated circuit has been designed, fabricated, and successfully demonstrated on actual sounds.

Corresponding author:

John G. Harris
Department of Electrical Engineering
University of Florida
453 EB, Dept. of Electrical and Computer Engineering
Gainesville FL 32611-2024 Phone: (352) 392-2652
harris@cnel.ufl.edu Fax : (352) 392-0044
1 Introduction

Human three-dimensional auditory localization has been regarded as a special case of passive localization by a system consisting of two sensors: the ears. In most human sound localization studies, the sound sources have been restricted to one of two planes: either the horizontal plane or the vertical (mid-sagittal) plane [11]. The distinction between horizontal and vertical localization also appears to be justified by differences in the principal spatial cues for horizontal and vertical localization (i.e. interaural difference cues vs. monaural cues). Sound waves incident onto the ears of a listener are reflected by the head on the side facing the incident wave, diffracted to the ear on the shadowed side of the head, and transmitted to the eardrum via the pinna. These reflections and diffractions produce interaural time differences (ITD) and interaural intensity differences (IID), which are well-known interaural difference cues.

For low-frequency components of sounds (below 1500Hz for humans), the phase-derived interaural time difference (ITD) can be used to localize the sound source. For these frequencies, the sound wavelength is at least several times larger than the head and the amount of shadowing (which depends on the wavelength of the sound compared with the dimensions of the head) is negligible. ITD localization is a well-studied system in biology (see e.g., [6]). The ITD cue has even been mapped to neuromorphic analog VLSI circuits with limited success on actual sound signals [2] [7].

Above 3000Hz, interaural phase differences become ambiguous by multiples of 360° and are no longer viable localization cues. For these high frequencies, the wavelength of the sound is small enough that the sound amplitude is attenuated by the head. The intensity difference of the log magnitudes at the ears provides a unique interaural intensity difference (IID) that can be used to localize. In the range of 1500 to 3000Hz, stimuli are too high in frequency to provide a usable phase cue and too long in wavelength to provide adequate IID. It is not surprising that localization performance is worst for sound signals in the range of
1500-3000Hz.

To the extent that the head and the ears are symmetrical, a stimulus presented at any location on the median plane should produce no interaural differences and thus, interaural differences should provide no cue to the vertical localization. Many studies have shown that when one ear is completely blocked, humans can still localize sounds in space, albeit at a worse resolution in the horizontal direction [3] [15]. Monaural localization requires that information is somehow extracted from the direction-dependent effects of the reflections and diffractions of sound off of the external ear (pinna), head, shoulder, and torso. The so-called “Head Related Transfer Function” (HRTF) is the effective direction-dependent transfer function that is applied to the incoming sound to produce the sound in the middle ear. Section 2 presents a brief introduction to sound localization cues. Section 3 discusses the hardware implementation of the single microphone sound localizer and its results. Section 4 concludes this work.

2 Sound localization cues

It is well known that sounds incident in the horizontal plane can be localized on the basis of phase-derived interaural time differences. It can be shown that at low frequencies the interaural time difference (ITD) can be approximated by: [15]

$$ITD \approx \frac{3a}{c} \sin(\theta_{inc})$$  \hspace{1cm} (1)

where $c$ is the ambient speed of sound, $a$ is the radius of the head, and $\theta_{inc}$ is the angle of incidence. At high frequencies, the primary localization cue in the horizontal plane is interaural intensity difference (IID). The IID is defined as

$$IID(dB) = 20 \log P_L - 20 \log P_R$$  \hspace{1cm} (2)

where $P_L$ and $P_R$ are pressure magnitudes at the left and right ears.
Batteau [1] was one of the first to emphasize that the external ear, specifically the pinna, could be a source of spatial cues that might account for vertical localization. He measured the impulse response of an enlarged pinna replica and concluded that the physical structure of the external ear introduced two significant echoes in addition to the original sound. One echo varies with the azimuthal position of the sound source, having a latency in the 0 to 80μs range, while the other varies with elevation in the 100μs to 300μs range. Batteau’s model is depicted in Figure 1. The output \( y(t) \) is related to the input \( x(t) \) as

\[
y(t) = x(t) + a_1 x(t - \tau_a) + a_2 x(t - \tau_v)
\]

where \( \tau_a, \tau_v \) refer to azimuth and elevation echoes respectively; \( a_1, a_2 \) are two reflection constants. Watkins [13] used Batteau’s results to synthesize direction-encoded sound, and he proved the model effective. Hiranaka et al. measured the impulse response of the pinna from nine subjects. Their results suggest that the human pinna works as a compound sound reflector which produces major reflected components within 350μs [5]. One of their conclusions from their measurements was: “The time interval between the first peak and those major reflections increases as the source moves downward.” The time interval provides the key to determine a sound source’s elevation angle. A convenient and mathematically tractable geometry for simulating the pinna’s curved surfaces is the parabola. The pinna can be approximately simulated if the concha wall and helix are combined to form one effective parabola with the ear canal is placed off the parabola’s axis, this can approximately simulate the pinna.

3 Hardware implementation and results

Our goal is to design a single microphone system for localizing a sound source in the vertical direction. As shown in Figure 2(a), a special reflection surface encodes the sound source’s direction (similar to the function of the external ear), a silicon cochlea functions as a band-pass filter bank, onset detecting circuitry detects and amplifies the energy change in the
sound signal, pulse generating circuitry transfers analog sound signals into pulse signals based on comparing sound signal with an adaptive threshold value, and delay time computation circuitry computes the echo’s time delay then decodes the sound source’s direction. Our implementation is based on biological principles, and our analog circuitry computes the sound source direction in real time.

3.1 Single microphone recordings

As discussed in the previous section, the external ear produces the interference pattern which encodes the sound source’s location in the front half plane. We may use any shape of reflection surface as long as the reflection echo caused by the surface provides one-to-one mapping between the echo’s delay time and the source’s direction. Thus, we propose two flat surfaces to compose the reflection structure in our model. The setup of our recording is depicted in Figure 2(b). The two flat surfaces are $S_1$ and $S_2$. A microphone is placed at distances $a_1$ and $a_2$ from the two flat surfaces, $d$ is the distance between the microphone and the sound source moving line (the dotted line in Figure 2(b)). As shown in Figure 2(b), a sound source is at $\angle \phi$ position. If the source is far enough from the reflection surface, the ray diagram is valid to analyze the sound’s behavior. If $C$ is the reflection point on the reflection surface $S_1$, and $\alpha$ is the incident and reflection angles, the direct and the reflected sound signals are analyzed based on the simple geometry. The direct and the reflected path can be expressed as:

$$d_1 = \frac{d}{\cos \phi} \quad (4)$$

$$R = r_1 + r_2 = \frac{a_1 + d}{\cos \alpha} + \frac{a_1}{\cos \alpha} \quad (5)$$

where $d_1$ and $R$ represent the direct path and the reflected path. The relation between the incident angle $\alpha$ and the source direction $\phi$ is

$$\alpha = \tan^{-1} \left( \frac{d \tan \phi}{2a_1 + d} \right) \quad (6)$$
with the above equations, the echo’s delay time is

\[ \tau = \frac{R - d_1}{c} \]  

(7)

where \( c \) is the speed of sound (approximately 340 m/s in air). As shown by the above geometric analysis, the echo’s delay time \( \tau \) decreases as the source position \( \phi \) moves from 0 to 90 degrees. A similar analysis can be made if the source moves in the opposite direction, and the reflection is caused by the other reflection surface \( S_2 \). Since the reflection path is longer for reflection surface \( S_2 \) than for reflection surface \( S_1 \); the echo’s delay time can be separated into two ranges. Therefore, the echo’s delay time encodes the source’s directions in a one-to-one mapping relation. Compared to the human external ear, our simple flat surfaces provide a convenient means to encode the sound source’s direction. The advantage of our flat reflection surface is that it is straightforward and simple, but its disadvantage is that it lacks sufficient flexibility to encode the source’s direction in more than a 1-D plane. However, this disadvantage can be overcome by replacing the flat surfaces with more complicated surfaces that provide more reflections to encode source directions. The trade-off of our proposed reflection surface is between simplicity and flexibility.

It has long been realized that the human cochlea separates the sound frequency components along its length. The cochlea transforms mechanical sound signals into electrical neuron pulses as the signals processed through the inner hair cells. Therefore, the cochlea is vital for handling daily audio signals such as speech. The silicon cochlea has been implemented successfully in analog VLSI circuitry [9] [14]; however, studying the cochlea is not the focus of this paper. To further simplify and test our proposed localization model, only impulse signals were tested. Therefore, the silicon cochlea is not necessary for initial testing of our localization model. To deal with more complicated signals (e.g. speech), a silicon cochlea implementation is required.

In our test setup, an Earthworks M30 microphone and a Lab1 amplifier [4] were used to record the impulse signal. The microphone output is fed into our system input.
3.2 Onset detection

Inspired by edge detection algorithms in visual processing, onset detection is used to segment sound [12]. The detection of an onset is produced by first taking the difference between two first-order low-pass filters. The onset detection is described as [12]

\[ O(t, k, r) = \int_0^t f_x(t - x, k)s(x)dx - \int_0^t f_x(t - x, k/r)s(x)dx \]  

where \( r > 1 \), \( k \) is some time constant, \( s(x) \) is the input sound signal, and \( f_x(x, k) = k \exp(-kx) \).

A hardware implementation of the above equation is depicted in Figure 3. In our model, sound signals from the special reflection surface microphone are fed into two low-pass filters which have different time constants determined by two bias voltages \( V_{on1} \) and \( V_{on2} \). The bias voltage \( V_{on3} \) determines the amplification of the difference. The output of the onset detecting circuit is \( V_{onout} \). This circuitry was designed for subthreshold CMOS operation because of low-power considerations and because time constants can be changed by several orders of magnitude in this region. The onset detection circuit determines significant increases in the signal energy and therefore segments sound events. By computing the delay time between two sound events (direct sound and its echo caused by the reflection surface), the system is able to decode the source's direction that was originally encoded by the reflection surface. In this impulse sound signal, it is much simpler if each sound event is transformed into a fixed width pulse. The delay time is computed by implementing autocorrelators. To achieve the fixed width pulse for each sound event, we may need to set a proper threshold to discriminate the sound events from noise.

3.3 Pulse generating circuit

The pulse generating circuit is depicted in Figure 4. The pulse generating circuit includes a self-resetting neuron circuit [10]. This self-resetting neuron circuit controls the pulse duration based on the bias voltage \( V_{neu3} \). As in our earlier discussion, an appropriate threshold is
required to discriminate sound events from noise. One input of the pulse generating circuit is the output of the onset detecting signal, $V_{on_{out}}$. Since the onset output is somewhat insensitive to noise and amplifies energy change, $V_{thresh}$ is set properly in the pulse generating circuit to generate a fixed width pulse as $V_{on_{out}}$ exceeds $V_{thresh}$. The other problem of the pulse generating circuit is the unwanted sound events that may confuse the system. These unwanted sound events are caused by environment reflections, such as the reflections from the desks, walls. These reflections are common in every day environments. From the geometric analysis of our reflection surface, a certain range of the echo's delay time can be expected. We take advantage of the expected values, take the echos only from the expected time range, and get rid of the unwanted reflections. This can be achieved with an adaptive threshold.

### 3.4 Adaptive threshold circuit

In order to cancel unwanted signals, we need to design an inhibition mechanism that suppresses signals arriving to our system outside of the expected time range. This inhibition is implemented in Figure 5. As the pulse generating circuit detects the first sound event (which is the direct sound signal), the threshold becomes high in a certain period of time to suppress the detection of the unwanted reflections (not from our reflection surfaces). The input of the adaptive threshold circuit is $V_{neu_{out}}$ which is the output of the pulse generating circuit. The output of the threshold circuit is $V_{thresh}$ which is input to the pulse generating circuit. When the pulse generating circuit detects a sound event, $V_{neu_{out}}$ rises, which changes $V_{thresh}$ from $V_{ref2}$ to $V_{ref1}$ as shown in Figure 5. The higher $V_{thresh}$ suppresses the detection. The suppression time is simply determined by the other self-resetting neuron circuit.
3.5 Time delay computation

The nervous system needs to perform a running autocorrelation analysis. The basic neural connections are shown in Figure 6 [8]. $A(t)$ is the input neuron, $A(t-\tau), A(t-2\tau), \ldots A(t-m\tau)$ is a delay chain. The original signal and the delayed signal are multiplied when $A(t)$ and $A(t-k\tau)$ feed $C_k$. Assuming the state of neuron A is $N_A(t)$. If each synaptic delay in the chain is $\tau$, the chain gives us $N_A(t)$ under various delays. $C_k$ fires simultaneously when both $A(t)$ and $A(t-k\tau)$ fire. Neuron $C_k$ connects neuron $D_k$. Excitation is built up at $D_k$ by the charge and discharge of $C_k$. The excitation at $D_k$ is therefore

$$D_k(t) = N_{C_k}(t) = N_A(t)N_A(t-k\tau)$$

(9)

Viewing the arrangement of Figure 6 as a neuron autocorrelator, the time-varying excitation at $D_1, D_2, \ldots D_k$ provides a spatial representation of the autocorrelation function.

Our system includes an autocorrelator (as described above) to compute the delay between two signals. The localization resolution of this system depends on the delay time $\tau$, and the number of the correlators. As $\tau$ decreases, the localization resolution is improved provided there are enough correlators. In this paper, 30 unit delay taps, and 10 correlators have been implemented on chip. The outputs of the 10 correlators display the time difference between two sound events. The delay time decodes the source’s direction. Therefore, the 10 correlators also display spatial information of source’s direction in vertical direction.

3.6 Simulation results

HSPICE simulation results of this system are discussed next. Figure 7(a) shows the input sound signal which is an impulse signal recording in our lab (a typical student office environment). Figure 7(b) shows the output of the onset detector (labeled 61), the pulse generating output (labeled 12), and the adaptive threshold (labeled 11). As the onset output exceeds the threshold, the output of the pulse generating circuit becomes high. Simultaneously, the
high value of the generated pulse turns on the adaptive threshold circuit to increase the threshold voltage. The adaptive threshold voltage suppresses the unwanted reflection which can be seen right after the direct signal (we believe the unwanted reflection is caused by the table). Figure 7(b) shows that the two sound events are detected successfully. The pulse signals are fed to the correlator circuit (as signal \( A \) in Figure 6). Figure 8(a) shows the two pulses which represent the two sound events (direct and its delayed reflection) and their delayed versions (as signal \( A(t - kr) \) in Figure 6). Figure 8(b) shows the correlator outputs (as signal \( D_k \) in Figure 6). Only one correlator (labeled 1010) has an active response.

Five recordings of the impulse sound source moving along the 1-D plane have been tested in the simulation of our system. For each test, only one of the 10 correlators responded actively which shows our implemented system works well in simulation.

### 3.7 Measurement results

The single microphone sound localizer circuit has been designed and fabricated through the MOSIS 2μm N-well CMOS process. Impulse signals are used to test the fabricated localizer chip. In this section, the test measurements are presented.

Figure 9 depicts the block diagram of the test setup. The M30 microphone picks up the direct impulse signal and echoes from our special reflection surface. The composite signals are fed into the input of the sound localizer after amplification. Our sound localizer chip receives the composite signal, computes the echo time delay, and sends out the localization result to a display circuit. The display circuit is composed of 4 LEDs with each LED representing a specific sound source location\(^1\). The sound localizer sends the computational result to turn on a specific LED signifying the echo time delay and consequently, the source direction.

In the test, the M30 microphone and the reflection surface are placed at fixed locations. The

---

\(^1\)Only four of the ten correlator outputs were used in this experiment.
speaker is moved along the dotted line shown in Figure 9. The M30 microphone is distance $d_1$ from the reflection surface and $a_1$ from the speaker moving line. Distance $d_1$ is 33cm and $a_1$ is 24cm in our test. The speaker's location is defined as $d_2$ as depicted in Figure 9. The test result is shown in Figure 10. Figure 10(a) shows the theoretical echo delay at various speaker locations. Figure 10(b) is the measurement of the setup depicted in Figure 9. The y-axis indicates LED 1 through LED 4. The x-axis represents the distance between the speaker's location ($d_2$ in Figure 9). The solid horizontal line in Figure 10(b) represents the theoretical results for which LED should respond for each displacement. The results show that localization is accurate within each region with possibilities of two LEDs responding in the overlap regions.

For the same echo delay time constant, the locus of curves of the sound source locations in our test setup (depicted in Figure 9) are shown in Figure 11. Figure 11 shows the sound source locations which have the same echo delay time. The symbol "o" represents the microphone location which is referenced as (0,0) in the plot. As the sound source is moved far away, each sound source direction ($\phi$ in Figure 9) has a certain echo delay time. Therefore, this single microphone sound localizer also localizes sound sources directions.

4 Conclusion

We have developed the first monaural sound localization system. This system provides a real-time model for human sound localization and has potential use in such applications as low-cost teleconferencing. More work is needed to further develop the system. We need to characterize the accuracy of our system and to test more interesting sound signals, such as speech. Our flat reflection surface is straightforward and simple, but it lacks sufficient flexibility to encode the source's direction in more than a 1-D plane. We plan to replace the flat surfaces with a more complicated surface to provide more reflections to encode a richer set of source directions.
Acknowledgments

This work was supported by an ONR contract #N00014-94-1-0858 and an NSF CAREER award #MIP-9502307. We gratefully acknowledge MOSIS chip fabrication and Earthworks Inc. for loaning the M30 microphone and amplifier.

References


Figure 1: Batteau external ear model

Figure 2: (a) Proposed localization model is inspired from the biological model (b) Special reflection surface to serve the role of the pinna
Figure 3: Sound signal's onset is detected by taking the difference of two low-pass filters with different time constants

Figure 4: Pulse generating circuit

Figure 5: Adaptive threshold circuit used to remove unwanted reflections.
Figure 6: Neural signal processing model

Figure 7: (a) The input sound signal: impulse signal recorded in typical office environment (b) HSPICE simulation of the output of the detecting onset circuit (label 61), the output of the pulse generating circuit (label 12), and the adaptive threshold circuit response (label 11)
Figure 8: (a) HSPICE simulations of the neuron signal denoted by $A(t - k\tau)$ in Figure 6 (b). The system response which shows only one output (signal $D_k$ in Figure 6) is active.

Figure 9: Block diagram of the test setup
Figure 10: Sound localizer chip test result

Figure 11: Echo time delay locus curve