Learning Sound Location from a Single Microphone

Ashutosh Saxena and Andrew Y. Ng
Computer Science Department,
Stanford University, Stanford, CA 94305, USA.
{asaxena,ang}cs.stanford.edu

***Abstract*— We consider the problem of estimating the incident angle of a sound, using only a single microphone. The ability to perform monaural (single-ear) localization is important to many animals; indeed, monaural cues are also the primary method by which humans decide if a sound comes from the front or back, as well as estimate its elevation. Such monaural localization is made possible by the structure of the pinna (outer ear), which modifies sound in a way that is dependent on its incident angle. In this paper, we propose a machine learning approach to monaural localization, using only a single microphone and an "artificial pinna" (that distorts sound in a direction-dependent way). Our approach models the typical distribution of natural and artificial sounds, as well as the direction-dependent changes to sounds induced by the pinna. Our experimental results also show that the algorithm is able to fairly accurately localize a wide range of sounds, such as human speech, dog barking, waterfall, thunder, and so on. In contrast to microphone arrays, this approach also offers the potential of significantly more compact, as well as lower cost and power, devices for sounds localization.**

I. Introduction

The ability to perform sound localization—i.e., to estimate the direction of a sound source—is important to many biological organisms, where sound can serve as a warning of danger, or be used to locate prey [11]. Further, sound localization has many important engineering applications, ranging from estimating the position of a speaker, to automatically deciding where to steer a directional microphone (or beamformer) or camera.

Typically, sound localization in artificial systems is performed by using two (or more) microphones. By using the difference of arrival times of a sound at the two microphones, one can mathematically estimate the direction of the sound source. However, the accuracy which which an array of microphones can localize a sound (using interaural time difference) is fundamentally limited by the physical size of the array [6], [5]. If the array is too small, then the microphones are spaced too closely together so that they all record essentially the same sound (with interaural time differences near zero), making it extremely difficult to estimate the orientation. Thus, it is not uncommon for microphone arrays to range from 10's of centimeters in length (for desktop applications) to many 10's of meters in length (for underwater localization). However, microphone arrays of this size then become impractical to use on small robots [24], [19]; even for large robots, such microphone arrays can be cumbersome to mount and to maneuver. In contrast, the ability to localize sound using a *single* microphone (which can be made extremely small) holds the potential of significantly more compact, as well as lower cost and power, devices for localization. Being able to do so is also an interesting and enlightening problem in its own right.

In biological organisms such as humans, "interaural time difference" is also used as a cue for sound localization; on organisms with two ears (or microphones), this is called the biaural time difference or the biaural cue. However, in humans the biaural cue cannot be used to estimate the elevation of the sound, nor can it be used to distinguish a sound coming from the front vs. one from the back. For example, if a sound source is directly in front of us, then the interaural time difference will be zero, regardless of the source's height. This cue in isolation also suffers from front/back ambiguity.

Humans can localize the full 3d direction of a sound with reasonably high accuracy, including both elevation and whether it is from the front or back. Indeed, humans can perform this task using even a *single* ear; this is known as monaural localization. They can perform this task because the sound measured in the inner ear changes as a function of the source's direction. Specifically, reflections from the ear pinna (the outer ear part of ear, also called the auricle) and the head changes the perceived sound in a way that is dependent on its source's direction. This allows humans and other organisms to perform monaural localization (including estimating the elevation of a sound source).

Monaural localization, however, is a challenging problem for artificial systems, because it requires prior knowledge of the possible sounds. Indeed, the ability of humans to estimate the direction of a sound monaurally is contingent on their familiarity with it. Specifically, even though sounds are modified by the ear depending on its incident angle, we note that in a narrow mathematical sense, it is actually impossible to determine a sound's direction from a monaural recording alone, because it is impossible to know whether a sound appears different because it is coming from a certain direction (and thus modified in a certain way by the pinna), or if it was originally like that. However, typical sounds found in our environments (and in natural environments) are not random—they have certain structure. Thus, it is by using our prior knowledge (perhaps gained through our years of experience with sound) about what sounds are likely that we

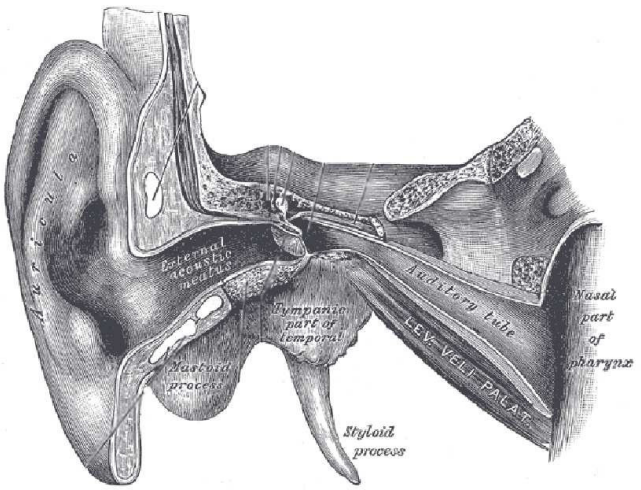
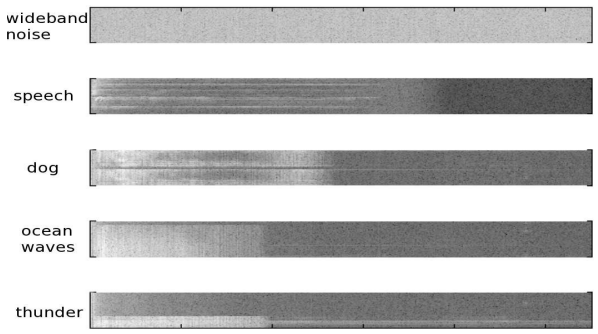


Fig. 1: Diagram of human ear. The outer ear (pinna) changes the spectral response of the sound depending on the direction it was coming from. (Image credit: [12])



(a) Spectrogram of various sounds

Fig. 2: (a) Spectrogram $\log S_t^2(f)$ of some typical sounds.

can estimate its most likely direction. Psychological studies have also found that monaural localization is somewhat poor in infants as compared to adults; one of many possible explanations for this could be that they have not yet had sufficient exposure to natural sounds to have built a good prior for what typical sounds are like [22].

In this paper, we will describe a machine learning approach to estimating the direction of a sound source, given only a single microphone (at a fixed position) and an "artificial pinna." We will present a generative model and evaluate it extensively with single microphone sound localization with different pinna designs on a variety sound sources, including noise, human speech, dog barking, and many natural sounds.

II. Monaural Cues: Biological Motivation

The sound appears different depending on what direction it is coming from, because of the pinna (Fig. 1). The pinna's convoluted structure causes sound waves, as they are gathered and funneled toward the ear canal, to experience overlapping, cancellation and reverberation influences. These influences change the spectral shape of the sounds in a direction-dependent way. This change provides the monaural information necessary for determining where a sound is coming from. However, this brings forth a curious fact: for one to perceive a changing sound as coming from the same source and maintaining a consistent acoustical quality, one must have a baseline familiarity with it. Indeed, [34] reports that our experience of a sound and our ability to locate it is contingent upon our familiarity with it. They suggest that we are attuned to the acoustic patterns as they are created by our own pinna, and that we must be familiar with a sound to begin with so that we know that it is being modified [11]. In other words, we must have a reference pattern of the sound from which to evaluate its acoustical qualities as it is experienced in different locations in space.

Even beyond the effects of the pinna, in humans there are other direction-dependent effects such as sounds reflecting off one's shoulder, and sounds being attenuated through passing around the head to reach the distal ear. In our approach (described in Section 3), we will model the overall effect of these direction-dependent modifications using a direction-dependent transfer function h_θ.

In the psychology literature, it is somewhat agreed on that binaural cues in humans are used primarily for determining azimuth (i.e., left-right direction) while monaural cues are, for the most part, used for determining elevation (up-down direction) [7], [20], [37]. Some researchers [8], however, believe that monaural cues are more important than binaural cues even for determining azimuth. As discussed earlier, when a sound is presented on the medial vertical plane (i.e., equidistant from the two ears), binaural differential hearing is impossible. Therefore, the determination of where a sound is located on that plane relies upon monaural pinna cues. [14]

In biological systems, estimating the location of sound source, and understanding the content of the sound, often represent two conflicting requirements. I.e., to understand the content of the sound, one would want to have minimum distortion; however, to localize the sound, one would want the sound to be different for different incidence angles. Different species have evolved ears to meet these requirements. For example, bats and owls [16], [21] which have to perform echo-localization have ears that affect the sound in a highly direction-dependent way, while the ears of pigs have only a mild direction-dependency.

III. Model

Here, we will formally describe our model. Let $s(t)$ be the sound source, and let $h_\theta(t)$ be the direction dependent transfer function (analogous to the direction dependent transfer function of the ear pinna). Then, the signal recorded by the microphone $y(t)$ is given as

$$y(t) = h_\theta(t) * s(t) + w(t) \tag{1}$$

where "$*$" denotes convolution, and $w(t) \in \mathcal{N}(0, \sigma_w^2)$ is additive white Gaussian noise.

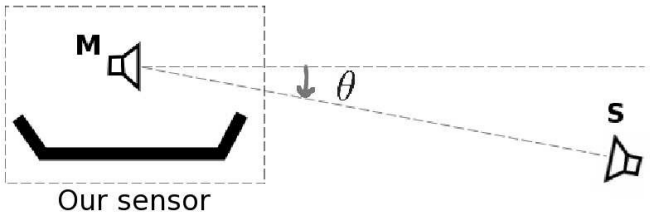


Fig. 3: Illustration of our recording system, showing the microphone (M), the speaker (S), and the angle (θ).

The direction-dependent response h_θ is time-invariant,[1] and we have approximated the microphone-pinna system as a linear system. Thus, we can represent $h_\theta(t)$ in the Fourier (frequency) domain [23] as $H_\theta(f) = \mathcal{F}[h_\theta(t)]$. Informally, the Fourier transform $\mathcal{F}[\cdot]$ gives a representation of $h_\theta(t)$ in terms of the amount by which it amplifies or attenuates the audio source at different frequencies f. With some abuse of notation, we will sometimes also write $H_\theta(f)$ as a vector H_θ. Finally, we applied standard noise excitation methods [18] to estimate H_θ for different directions θ, on our specific microphone-pinna setup (described in Section IV). This thus gives us a fairly accurate model of our system's direction dependent response to different audio sources.

We now describe our model for the audio source $s(t)$. Various properties of typical sounds $s(t)$ change as function of time. For example, when we speak a word such as "hello," the acoustics are extremely different for the "h", "e", "l" and "oo" parts of the word. A similar effect holds true to other natural sounds as well. To capture these time-dependent effects, we will apply the Short-term-Fourier-Transforms (STFT) (with a Hann window centered at time t, see [23] for details) to compute the frequency-domain representation $S_t(f)$ of the signal at time t. The squared magnitude of this representation, $S_t^2(f)$, is called the spectrogram. We will also write S_t^2 to denote a vector containing all of $S_t^2(f)$'s components.

Even though the sound signal $S_t^2(f)$ is stochastic and its (first and second order) statistics change as a function of time, many sounds in nature can be modeled as having time-invariant statistics when we consider only short time windows. E.g., each of the parts of the "hello" signal (each phoneme) can be accurately modeled as being wide-sense stationary (i.e., for a phoneme the mean and covariance remain relatively constant), even though the entire word as a whole is not. Following common practice in audio signal processing [26], [32], we will therefore model the $S_t^2(f)$ using a Hidden Markov Model (HMM) [27]. We use an observation model $P(S_t^2|z_t) = \mathcal{N}(S_t^2; \mu_{z_t}, \Sigma_{z_t})$, where $z_t \in \{1, ..., K\}$ is the state of the HMM at time t. The HMM transitions from one state z_t to another state z_{t+1} according to the model's state transition probabilities $P(z_{t+1}|z_t)$. We trained an HMM model of typical sounds using the standard EM (Baum-Welch), with the parameters initialized by using K-means clustering to group the observations S_t in our training set into different discrete states z_t. The model was trained using a variety of natural and artificial sounds (see Section IV-A for details).

We can now express Eq. 1 in the frequency domain, where the convolution operation becomes multiplication:

$$Y_t(f) = H_\theta(f)S_t(f) + W_t(f) \tag{2}$$

Here, W_t is the Gaussian noise w expressed in the frequency domain. This equation also sheds light on why monaural localization is challenging, and why the direction of a sound source is ambiguous given only one microphone recording. Specifically, we have a noisy observation of the product Y_t of two unknown quantities H_θ (because θ is unknown) and S_t; and, our goal is to decompose the product back into the two original multiplicands. Clearly, there are many possible choices for H_θ and S_t that would give rise to the same product; thus, monaural localization in an inherently ambiguous problem.

Fortunately, only certain sound signals S_t are likely. Thus, using a model of typical sounds, we can attempt to find the most probable decomposition of Y_t into H_θ and S_t. For example, S_t tends to be correlated in the frequency-domain; e.g., if its value at a particular frequency is high, then the values at the neighboring frequencies are also likely to be higher. Our learned HMM model attempts to capture such properties of typical sounds.

To infer the value of θ, we begin by computing the power of the signal:

$$Y_t^2(f) = H_\theta^2(f)S_t^2(f) + 2W_t(f)H_\theta(f)S_t(f) + W_t^2(f) \tag{3}$$

We will compute the first and second-order moments of $Y_t^2(f)$. We have that $\Sigma_{z_t}(f_1, f_2)$ is the covariance between $S_t^2(f_1)$ and $S_t^2(f_2)$ in the HMM's observation probability given state z_t. Let $\sigma_{z_t}^2(f) = \Sigma_{z_t}(f, f)$ be the diagonal elements of these covariance matrices. We now have:

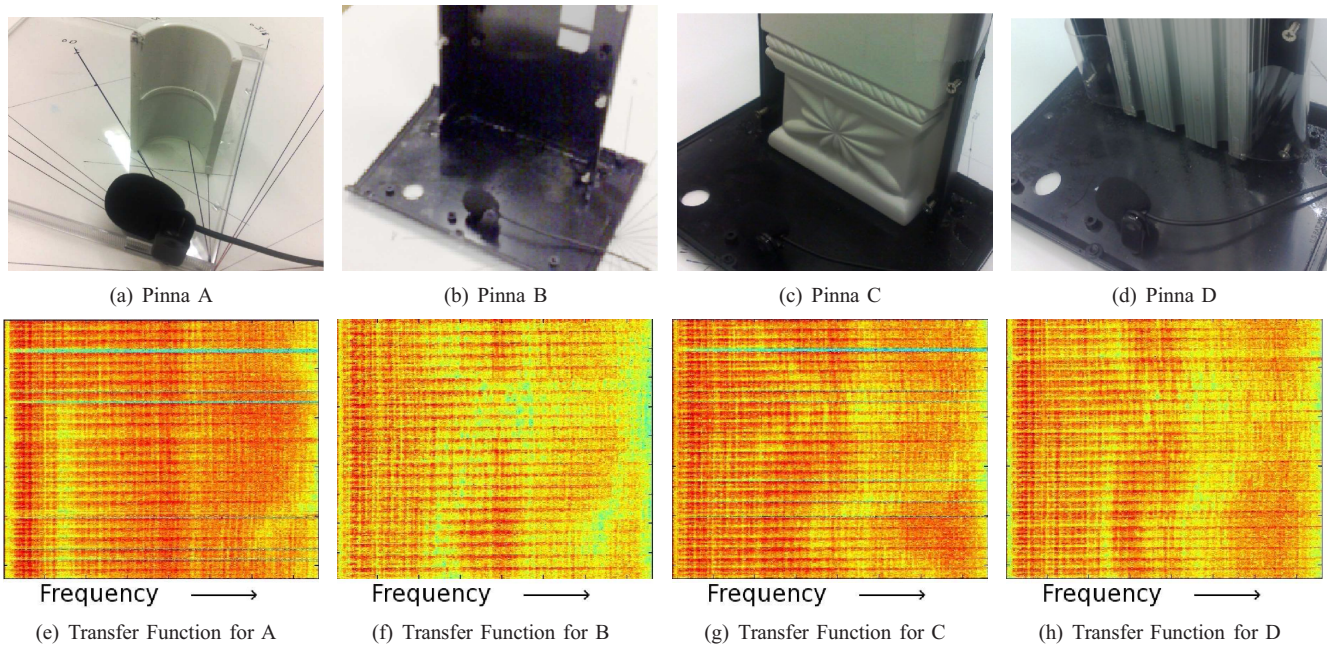
$$\begin{aligned} E[Y_t^2(f)|z_t] &= H_\theta^2(f)\mu_{z_t}(f) + \sigma_w^2(f) \\ \mathrm{Var}[Y_t^2(f)|z_t] &= H_\theta^4(f)\sigma_{z_t}^2(f) + \\ &\quad 4H_\theta^2(f)\mu_{z_t}(f)\sigma_w^2(f) + 2\sigma_w^4(f) \\ \mathrm{Cov}[Y_t^2(f_1), Y_t^2(f_2)|z_t] &= H_\theta^2(f_1)H_\theta^2(f_2)\Sigma_{z_t}(f_1, f_2), f_1 \neq f_2 \end{aligned} \tag{4}$$

Having computed the moments of Y_t^2 (conditioned on z_t), we will approximate $P(Y_t^2|z_t, \theta)$ as a Gaussian $\mathcal{N}(E(Y_t^2|z_t), \mathrm{Cov}(Y_t^2|z_t))$ with the corresponding mean vector and covariance matrix.

Inference: Given an input Y_t, we would like to infer the most likely value for θ. We have:

$$\begin{aligned} \hat{\theta} &= \arg\max_\theta \log P(Y_1^2, \ldots, Y_T^2|\theta) \\ &= \arg\max_\theta \log \sum_{\{z_t\}} \left(\prod_t P(Y_t^2|z_t, \theta) \right) p(z_1) \prod_{t=2}^{T} p(z_t|z_{t-1}) \end{aligned} \tag{5}$$

[1] A time-invariant system is one for which a time shift or delay of input sequence causes a corresponding shift in the output sequence.



(a) Pinna A (b) Pinna B (c) Pinna C (d) Pinna D

(e) Transfer Function for A (f) Transfer Function for B (g) Transfer Function for C (h) Transfer Function for D

Fig. 4: Pictures of our sensors—a single microphone with different "artificial pinnas." The bottom row shows their direction-related transfer function H_θ as a function of orientation (vertical axis). Pinnas that have very different response for different orientations are better for localization, e.g., pinna B and C. (Note the "wavy" effects.) Also note that H_θ has relatively small/subtle variations compared to the sound sources, making it challenging to infer which parts of a recorded signal were due to H_θ, and which parts due to the original acoustic signal. (Best viewed in color.)

$P(Y_t^2|z_t, \theta)$ is given by Eq. 4, and $p(z_t|z_{t-1})$ are given by the HMM state transition probabilities.

We perform the optimization for computing $\arg\max$ in two steps. First, we discretize θ in steps of 15° and compute the score for each of the values for $\theta \in \{0, 15, 30, \ldots, 345\}$, using the forward-backward algorithm [27] to compute the HMM observation probability. In the next step, we do the same inference in steps of 2 degrees in the local neighborhood of the best angle found.

IV. Experiments

A. Experimental Setup

In order to have a transfer function that depends on direction of the sound, we designed a set of small partial enclosures for the microphone, that serves as our "artificial pinna." They consist of a base plate with vertical walls of various shapes. The first one in Fig. 4a consists of a smooth pinna (made of a semi-cylindrical tube); the second one in Fig. 4b contains a few "corners" in the pinna but is otherwise smooth. In the third one (Fig. 4c), we use a plastic-cast that has smooth grooves built on it in various directions. Finally, in the fourth one in Fig. 4d, we use a surface that has sharp discontinuities in it.

B. Results

To train our model, we used white Gaussian noise recorded in our setup to estimate the direction related transfer function $H_\theta(f)$ at angles $\theta = (0, 15, 30, ..., 345)$ and the values were interpolated for finer angles. To train the HMM model for $P(S_t)$, we did not need to collect data by actually using our setup. We used 50 states in our HMM, and simply trained the HMM using various sounds available from publicly available datasets—human speech (also from [2] but from different speakers compared to the test set and in different environments), animal sounds (cat, horse, puma, elephant) and natural sounds (wind, river).

Our test data comprised a variety of different sound sources: human speech (with a range of accents, [2]), animal sounds such as a dog barking and growling, and other natural sounds such as ocean waves, thunderstorm, waterfall, etc. We recorded data by playing the sounds through a high-fidelity speaker, playing the speaker at 26 locations, chosen randomly in the range $[0^\circ, 360^\circ)$. A total of about 5 minutes of data was recorded at each angle θ. Our data was collected in a normal office environment, which had some background noise from sources such as air conditioning, computer fans, vehicles on the roads outside the building, etc. While the recordings were being made, sometimes there were people talking and walking in the corridors outside, there was also noise from people opening and shutting doors in nearby offices, and so on. The signal to noise ratio $10 \log_{10} P_s/P_n$ was about 18 dB for the the office environment. In addition, there was some room reverberation. Lastly, we also tested the algorithm on a pure tone. Note that these sounds are quite different from those in our training set.

In the various design choices for the pinna, Table I shows the comparative results. The pinna with a smooth surface

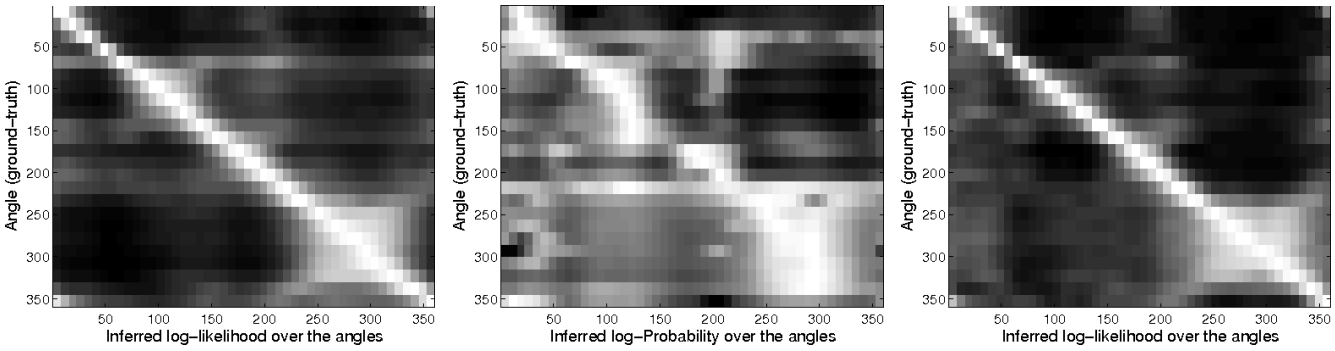


Fig. 5: Log-likelihood for predicting the angle. From left, noise, dog barking, and a segment of speech for Pinna C.

(Pinna A, Fig. 4a) the transfer function varies very smoothly as a function of the distance. I.e., the pinna does not modify the sound spectrum in a way that is unique to direction-dependency. The same change in transfer function could come because of a different sound source. Pinna B that has two sharp corners and also a few protruding notches produces better direction predictions. Pinna C has a number of smooth groves, and this produces a very unique spectral signature in the transfer function. In Fig. 4g, note a recurring pattern in the transfer function that changes as a function of the direction. Finally, we tested with pinna D that has a number of grooves and discontinuities; this however produced worse results. We believe this is because pinna D, made of metal, produces resonances in the columns.

Fig. 5 shows the log-likelihood estimated by our generative model for different values of θ and under different types of sounds. Our algorithm typically becomes confused when the angle is in the range from 235° to 345°. These angles correspond to the microphone in our model receives comparatively less reflections from the pinna (see Fig. 3). This results in the amplitude of main sound signal overwhelm the effects due to pinna. Better "pinna" designs that enclose the pinna would perhaps further increase the accuracy further.

Table I shows that for wide-band noise-like signals, the average error for pinna C was 4.3° (8.8° for B). Our error was somewhat higher (18.3° for C and 14.2° for B) for natural sounds with sharp onsets and transients (e.g., dog bark) that occur only for a very short duration in the test set. Because of the short time that these transients occur in, our algorithm is unable to accumulate sufficiently accurate statistics of the sound to make good predictions. However, for other "smoother" natural sounds that have less sharp transients, the errors are generally lower. For example, for ocean/waterfall sounds, it was 9.3° for C and 11.8° for B. The speech data was also particularly challenging, since it comprised multiple individuals speaking, often with their voices overlapping. Even then, our algorithm was able to predict directions with an error of 19.3° and 7.7° respectively. Overall, our pinna C works best with an overall average error of 13.5°. By way of comparison, [1] reports that monaural (single ear) localization error in humans ranges from 12° to 40°.[2]

We also tested our algorithm on pure tone (where all the acoustic energy is in one frequency component, such as the sound generated by a tuning fork), where there is no spectral shaping that happens because of the pinna. Thus, it is impossible to infer the direction of the sound. Indeed, psychological studies have shown that humans cannot localize pure tones [7], and our algorithm similarly fares poorly on it. (We believe that it performs slightly better than random because of the static noise in the speaker.)

V. Related Work

There is a long and distinguished history of work on sound localization from multiple microphones (e.g., microphone arrays). It is impossible to do justice to this literature, but [36], [15] offer detailed reviews. Examples of this work include Ben-Reuven and Singer [4], who used a discriminative classifier for binaural sound localization, and Thrun [35], who localized a set of microphones using unknown sounds emitted from unknown locations. Roman and Wang [29] used binaural cues to track multiple sources. They [28] also used inter-aural time and level differences to perform source separation. Fischer [10] used binaural cues and a head-related transfer function together to explain localization in owls.

Pearlmutter and Zador [25] used an overcomplete dictionary as their model of "natural sounds," and used L_1 sparse approximation to separate individual sounds from a mixture when the individual sounds were spatially separated. Zakarauskas and Cynader [38] used Head-Related-Transfer-Function (HRTFs) to localize broadband noise-like sounds; but their method would not apply to more generic sounds. Kristjansson et al. [17] separated sound from a male speaker and a female speaker recorded using a single microphone. There is also a large body of work in audio understanding applications other than localization, such as speech recognition and speech reconstruction [30], [3]. Chau and Duda [9] combined monaural and binaural cues for localizing noise-like sounds. Machine learning algorithms have also been

[2]Infants have worse accuracies in localizing sounds using monaural cues [22].

TABLE I: Average error for sound localization in degrees of the Generative model. In this test, the possible values for the angle was $\theta \in [0°, 360°)$.

PINNA	A	B	C	D
RANDOM ESTIMATE	90.0°	90.0°	90°	90°
WIDEBAND NOISE (E.G., RADIO STATIC)	42.6°	8.8°	**4.3°**	22.3°
MIXED SPEECH	67.8°	19.3°	**7.7°**	21.35°
NATURAL (DOG)	55.7°	**14.2°**	18.3°	60.28°
NATURAL (E.G., OCEAN, WATERFALL)	42.1°	11.8°	**9.3°**	42.7°
NATURAL (SHARP NOISES: THUNDER)	48.7°	21.6°	**18.8°**	39.6°
PURE TONE	88.7°	89.1°	86.4°	82.6°

developed for addressing other ill-posed problems such as inferring depth from a single still image [31], [33].

Harris, Pu and Principe [13] were also able to infer the direction of sound from a single microphone by built a hardware circuit to detect the difference in arrival time of a pulse (which is a signal that is non-zero only at one instant in time) and its reflection. However, their method only works for pulses and in relatively noise free environments.

VI. CONCLUSION

Monaural localization is an ability for many biological organisms, but represents a difficult estimation problem because the sound's incidence angle is intrinsically ambiguous—an acoustic signal may sound a certain way because it had arrived from a direction that caused it to be modified by the pinna in a specific way, or it may have originally sounded like that. The ability to localize sounds using a single microphone also holds the potential of much smaller (and less expensive) devices for sound localization than the large microphone arrays that are typically used today. In this paper, we presented a machine learning approach to monaural localization, using a single microphone and an artificial pinna. Our method models the prior distributions of sounds as well as the direction dependent transfer function of the pinna. These algorithms were able to estimate the orientation of a wide range of natural and artificial sounds fairly accurately.

ACKNOWLEDGMENTS

We thank Olga Russakovsky for useful discussions and Sai Soundararaj for help in collecting data.

REFERENCES

[1] Sound localization by human listeners. *Annu Rev Psychol*, 42, 1991.
[2] Announcing the ami meeting corpus. *ELRA Newsletter*, 11(1), 2006.
[3] K. Achan, S. T. Roweis, and B. J. Frey. Probabilistic inference of speech signals from phaseless spectrograms. In *NIPS*, 2003.
[4] E. Ben-Reuven and Y. Singer. Discriminative binaural sound localization. In *NIPS*, 2003.
[5] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. Springer, 2008.
[6] M. Brandstein and D. Ward. *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.
[7] R. Butler and D. Green. Sound localization by human listeners. *Perception and Psychophysics*, 51:182–186, 1991.
[8] S. Carlile. The auditory periphery of the ferret ii: The spectral trasnformations of the external ear and their implications for sound localization. *J Acoust Soc America*, 88, 1990.
[9] W. Chau and R. O. Duda. Combined monaural and binaural localization of sound sources. In *Signals, Systems and Computers*, 1995.
[10] B. J. Fischer. Optimal models of sound localization by barn owls. In *NIPS*, 2003.
[11] C. D. Geisler. *From sound to synapse*. Oxford University Press, 1998.
[12] H. Gray. *Henry Gray's Anatomy of the Human Body*. 1858. http://en.wikipedia.org/wiki/File:Gray907.png.
[13] J. G. Harris, C.-J. Pu, and J. C. Principe. A monaural cue sound localizer. *Analog Integrated Circuits and Signal Processing*, 23, 2000.
[14] A. F. Jahn. *From sound to synapse*. Singular Thomson Learning, 2001.
[15] F. Keyrouz, K. Diepold, and S. Keyrouz. High performance 3d sound localization for surveillance applications. In *Advanced Video and Signal Based Surveillance*, 2007.
[16] E. Knudsen. Instructed learning in the auditory localization pathway of the barn owl. *Nature*, 417, 2002.
[17] T. Kristjansson, H. Attias, and J. Hershey. Single microphone source separation using high resolution signal reconstruction. In *IEEE*, 2004.
[18] B. McCarthy. *Sound Systems: Design and Optimization*. Focal Press, 2006.
[19] J. Michels, A. Saxena, and A. Ng. High speed obstacle avoidance using monocular vision and reinforcement learning. In *ICML*, 2005.
[20] J. C. Middlebroks and D. M. Green. Sound localization by human listeners. *Annual Review of Psychology*, 42:135–159, 1991.
[21] A. Moiseff and M. Konishi. The owl's interaural pathway is not involved in sound localization. *Journal Comp Physiology A: Neuroethology, Sensory, Neural and Behavioral Physiology*, 144(3), 1981.
[22] B. A. Morrongiello. Infants' monaural localization of sounds: Effects of unilateral ear infection. *J Acoust Soc America*, 86(2), 1989.
[23] A. V. Oppenheim and R. W. Schafer. *Discrete-Time Signal Processing*. Pearson Education, 1999.
[24] M. Park, S. Chitta, A. Teichman, and M. Yim. Automatic configuration recognition methods in modular robots. *IJRR*, 27(3-4):403–421, 2008.
[25] B. A. Pearlmutter and A. M. Zador. Monaural source separation using spectral cues. In *ICA*, 2004.
[26] J. Picone. Signal modeling techniques in speech recognition. *Proc of the IEEE*, 81(8):1215–1247, 1993.
[27] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc IEEE*, 77(2):257–286, 1989.
[28] N. Roman and D. Wang. A classification-based cocktail-party processor. In *NIPS*, 2004.
[29] N. Roman and D. Wang. Binaural tracking of multiple moving sources. *IEEE Tran Audio, Speech, and Language Processing*, 16(4), 2008.
[30] S. T. Roweis. One microphone source separation. In *NIPS*, 2000.
[31] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. In *NIPS*, 2005.
[32] A. Saxena and A. Singh. A microprocessor based speech recognizer for isolated hindi digits. In *IEEE ACE*, 2002.
[33] A. Saxena, M. Sun, and A. Y. Ng. Learning 3-d scene structure from a single still image. In *ICCV workshop on 3D Representation for Recognition (3dRR)*, 2007.
[34] J. Sodnik, S. Tomazic, R. Grasset, A. Duenser, and M. Billinghurst. Spatial sound localization in augmented reality environment. In *Proc Computer-Human Interaction special interest group (CHISIG)*, 1992.
[35] S. Thrun. Affine structure from sound. In *NIPS*, 2005.
[36] D. Wang and G. Brown. Computational auditory scene analysis: Principles, algorithms and applications. *IEEE Tran Neural Networks*, 19(1), 2008.
[37] F. L. Wightman and D. J. Kistler. Monaural sound localization revisited. *J Acoust Soc America*, 101(2):1050–1063, 1997.
[38] P. Zakarauskas and M. Cynader. A computational theory of spectral cue localization. *J Acoust Soc Am*, 94(3), 1993.