MACHINE LEARNING FOR SOUND SOURCE ELEVATION DETECTION

Hugh O'Dwyer, Enda Bates and Francis M. Boland

Department of Electrical and Electronic Engineering
Trinity College Dublin
odwyerh@tcd.ie

ABSTRACT

Recent studies have shown the effectiveness of using Machine Learning (ML) to perform Sound Source Localization (SSL) tasks. However, most of this research has been primarily aimed at performing SSL in the azimuthal plane only. While the Interaural Cues which allow us to perceive azimuthal source location are well researched, cues which allow us to determine elevation location are understood to a lesser degree. It is generally regarded that spectral cues resulting from reflections of sound from the head, shoulders and pinnae are responsible for our perception of elevated sound. In this study we use Machine Learning to determine the proficiency of a variety of Interaural and Spectral cues in determining sound source elevation.

1. INTRODUCTION

Sound Source Localization is a task that is constantly performed by humans to a high level of proficiency. For instance, even in noisy environments we can successfully locate multiple speakers occurring simultaneously. This is part of what is known as the cocktail party effect whereby individuals can selectively choose to listen to one speaker while ignoring another. Understanding how we perform SSL in both planes could have applications in Augmented Reality (AR), Virtual Reality (VR) as well as in communication devices.

1.1. Interaural Cues

The cues which help us to determine the azimuthal location of a source are regarded to be interaural, i.e. relating to both ears. The most prolific of these cues are Interaural Time Difference (ITD) and Interaural Level Difference (ILD) [1, 2]. ITD is the difference in time of arrival of a signal between two ears and ILD is the difference in signal level [3]. While ITD and ILD are good cues for asserting azimuthal location they do not provide much information in determining elevation. This is particularly true towards the front and back of a listener as ITD and ILD vary only slightly for different elevations [4]. This can be seen in Figure 1. Because of this it is important to examine other cues and their effectiveness at determining sound source elevation.

The Cross-Correlation Function (CCF) is a measure of the displacement of one signal relative to the other. Using a sliding window the CCF measures the similarity of two signals as a function of time with a peak in the CCF at the point (or lag) at which the two signals match. As demonstrated in [5], CCF provides a rich amount of information for sound source localization as the variation in the sidelobes of a CCF vary with source azimuth as well as elevation.

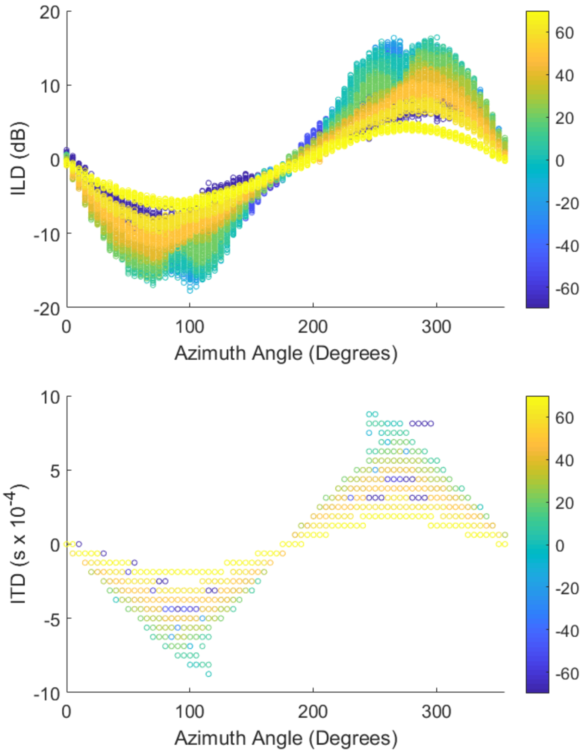

Figure 1: The variation of ILD and ITD across different azimuth and elevation angles. Here elevation is given according to the colourbar.

1.2. Monaural Cues

Elevation detection has been shown to rely on frequency dependent cues such as the reflection of sound waves off the ears, shoulders and torso [6]. These reflections vary with elevation and are responsible for spectral peaks and notches in an individuals Head Related Transfer Function (HRTF). These spectral changes occur in the frequency range above

5kHz [7]. Studies in the 1990s [8, 9] showed the feasibility in estimating elevation from cochlear processed signals. The processing of sound by the cochlea in the inner ear can be replicated using Gammatone Filters. Gammatone Filters separate sound into an array of overlapping frequency bands similarly to how sound is processed by the cochlea. Using ML, Youssef et al [2], presented promising results for elevation estimation using Gammatone Filter energies [GFEs]. The energies from these bandpassed signals capture the reflections caused by anthropometric features at low and high frequencies.

Many studies employ Mel-Frequency Cepstral Coefficients (MFCCs) when performing ML in audio systems. They are a feature which perform particularly well for tasks involving speech. A particular study demonstrates their proficiency in detecting speech in noisy and reverberant environments [10]. However, the potential of MFCCs in source direction detection has not received significant attention. The Mel-Frequency Cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCCs are coefficients which collectively make up the Mel-Frequency Cepstrum.

1.3. Machine Learning and Sound Source Localization

Until recently, localization models were implemented using Gaussian Mixture Models (GMMs) and lookup-table based hearing models. One such model presented by Ashby et al. [11], demonstrated a high level of accuracy in predicting both the azimuth and elevation of a sound source using ITD and ILD calculations measured across 4 pairs of microphones mounted spherically on a neck and torso simulator. Ma et al. [5], report a high level of accuracy in determining source azimuth to within 4° using a machine learning based approach. This localization study was performed using spatialized speech signals corrupted with noise to increase robustness. Both of the aforementioned studies incorporated a means of performing head rotation into their predictions. The purpose of this is to eliminate front-back confusion. Research presented by Lovedee and Murphy [12], used Deep Neural Networks (DNNs) to predict source azimuth using Head Related Impulse Responses (HRIRs). Although this method does not achieve the same level of accuracy as [5], it does present promising findings pertaining to the use of DNNs and sound source localization.

The research presented in this paper investigates how ML can be employed to determine sound source elevation in anechoic and reverberant environments. Section 2 of this paper describes the methods used to generate the feature sets used to train our Machine Learning algorithm. The implementation and architecture of our algorithm is also discussed in this section. Section 3 presents the results of several tests performed using different test stimuli and input feature sets. The results of these tests are discussed in Section 4 while Section 5 concludes with remarks on the outcomes of the study and future applications of this research.

2. METHODS

This study can be broken down into three main components; The creation of a database of stimuli, the extraction of features from these stimuli and the use of these features in training ML algorithms to detect sound source elevation. Each of these components are discussed in the proceeding subsections.

2.1. Stimuli Creation

Speech signals from the TIMIT database [13] were spatialized by convolving them with pairs of HRIRs from the SADIE database [14]. These HRIRs were measured on a KEMAR dummy head. Only HRIRs in the front hemisphere were used to eliminate front-back confusion from the study. In total, 555 HRIR pairs were used to generate our training stimuli. These were measured across the front hemisphere of the listener in increments of 5° in the azimuth and from -70° to +70° in elevation (in 10° increments). As in [5], the sampling rate used throughout this study was *fs* = 16kHz. HRIR signals were resampled to this rate. 5 random speech samples for each HRIR pair were selected from the TIMIT training database.

2.2. Feature Extraction

Feature extraction was performed using a signal processing front end in MATLAB. This process was performed using hamming windows of length 20ms with an overlap of 10ms. Values for each feature (ILD, ITD, CCF, MFCCs and GFEs) were then measured at each window.

ILD was measured using the full bandwidth signal such that,

$$ILD = 20log_{10}(E_l/E_r) \tag{1}$$

where E_l and E_r are the root mean square values of the signals in the left and right ears.

The Cross-Correlation Function between two signals can be calculated as follows,

$$CCF_{x_l x_r} = \sum_{n=0}^{N-1} x_l(n) x_r(n-k) \tag{2}$$

where x_l and x_r are the left and right signals, *N* is the length of both signals and *k* is the delay evaluated for each point along the length of the signals. Each CCF used in this system was evaluated for a lag range of ±1ms from the centre point of lag value zero. As the sampling rate used in this study is 16kHz, this resulted in a 33 dimensional vector for each CCF. After calculating CCF, the ITD was calculated simply as the lag value at the peak of the CCF. This is

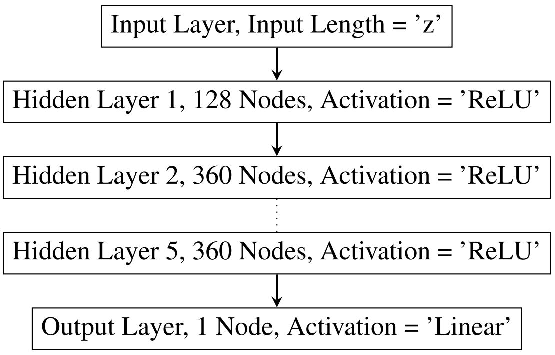
Input Layer, Input Length = 'z'
Hidden Layer 1, 128 Nodes, Activation = 'ReLU'
Hidden Layer 2, 360 Nodes, Activation = 'ReLU'
Hidden Layer 5, 360 Nodes, Activation = 'ReLU'
Output Layer, 1 Node, Activation = 'Linear'

Figure 2: The architecture of the elevation estimating Neural Network.

recognised as the difference in path length travelled by the signal arriving at both ears.

Gammatone Filter Energies (GFEs) and MFCCs are frequency dependent cues. A bank of 20 overlapping Gammatone Filters were used to analyse the binaural signals. These filters had centre frequencies uniformly placed along the equivalent rectangular bandwidth (ERB) scale between 80Hz and 8kHz. Due to the low sampling rate, higher frequencies could not be analysed. The energy of each band is computed individually for each ear by calculating the RMS value of the signal. A resulting vector of length 40 is generated. 13 MFCC values are calculated similarly for each signal resulting in a total of 26 MFCCs.

2.3. Machine Learning

Table 1: *Input vector length for each set of features examined.*

Feature Set	Input Length, 'z'
ITD	1
ILD	1
CCF	33
GFEs	40
MFCCs	26

Machine learning algorithms were designed to map each feature or combination of features to elevation angle. These algorithms were implemented using Keras as a high-level API to run a Tensorflow backend [15] [16]. A feedforward architecture was implemented, the topology of which connects the outputs in each layer to each node in the subsequent layer. The network consisted of an input layer, 5 hidden layers and an output layer as can be seen in Figure 2. The number of nodes in the input layer varied depending on the feature set being examined as shown in Table 1. Each feature was trained and tested separately to determine their effectiveness in predicting elevation. The first hidden layer contained 128 hidden nodes while the others consisted of 360 nodes. The Rectified Linear Unit (ReLU) activation function was used in each Hidden Layer. The output layer used a Linear activation function which uses regression to estimate the approximate elevation angle of the input feature set.

Each network was trained for a maximum of 250 epochs. A callback function was included to terminate training if the loss of the model showed no significant decrease over a period of 16 epochs. Mini-batches of size 60 were used to train the network. The Stochastic Gradient Descent (SGD) optimizer was used in training. Its learning rate was set to 8e^{-}6.

After training, our results are measured on the algorithms performance across 4 separate environments; each of which consists of unseen speech samples taken from the TIMIT Testing database [13]. In the anechoic environment these speech samples are simply spatialized in the same way as the training stimuli. For the 3 reverberant environments, additional reverb is added to the testing stimuli using Room Impulse Responses (RIRs) from the OpenAir Impulse Response library [17]. These RIRs are captured in real world environments, details of which can be seen in Table 2.

Table 2: *Binaural Room Impulses.*

	Description	Excitation Signal	RT_{60}
A	Living Room	Sine Sweep	0.2s
B	Church	Balloon Pop	0.53s
C	Mine Shaft	Sine Sweep	0.71s

3. RESULTS

Performance is measured under 3 headings, percentage of test stimuli correctly classified to within 5° of the actual elevation, those correctly classified to within 10°, and the Mean Square Error (MSE) of all the testing classifications. The results achieved by our NN in localizing sound source elevation in both anechoic and reverberant environments can be seen in Table 3. Multiple features were trained to determine how each feature contributed towards elevation detection. As described in [18], the just noticeable differences (JNDs) in human perception of elevation are greater than for azimuth which has a JND of approximately 4°. As such an accumulative prediction accuracy to within 10° is a good indication of performance by the system.

4. DISCUSSION

As expected, the interaural features ITD and ILD perfromed no better than random chance in predicting elevation. ILD did however perform marginally better than ITD indicating that elevation detection does rely more on frequency dependent cues than on time dependent cues. As noted in [19],

Table 3: *The performance each feature set in predicting sound source elevation in an anechoic environment and 3 separate reverberant environments.*

Features	Anechoic			Room A			Room B			Room C		
	± 5°	± 10°	MSE	± 5°	± 10°	MSE	± 5°	± 10°	MSE	± 5°	± 10°	MSE
ITD	6.7%	13.1%	36.96°	6.7%	13.33%	36.88°	6.67%	13.33%	36.87°	6.67%	13.32%	36.84°
ILD	7.31%	14.76%	35.5°	7.1%	14.46%	35.64°	7.2%	14.62%	35.61°	7.27%	14.69%	35.57°
CCF	29.54%	50.36%	16.38°	29.88%	50.08%	16.85°	29.95%	50.23%	16.77°	27.63%	47.3%	18.56°
GFEs	6.67%	13.33%	37.35°	6.67%	13.33%	37.35°	6.67%	13.33%	37.35°	6.67%	13.33%	37.35°
MFCCs	96%	99.05%	1.73°	91.85%	97.75%	2.26°	77.66%	84.41%	10.08°	59.85%	71.5%	14.02°
MFCCs/CCF	96.9%	99.2%	1.59°	94.38%	98.28%	2.03°	79.42%	84.75%	10.07°	62.5%	74.36%	12.97°

the CCF between two binaural signals proved to be a good predictor for elevation. This is exhibited by the fact that approximately 50% of testing stimuli were correctly identified to less than 10° for each of the 4 environments.

Unlike in [2], Gammatone Filter Energies proved to be a poor indicator of elevation, performing no better than chance. This may be due to the particular architecture behind the algorithm although such poor results indicate strongly that GFEs are fundamentally lacking for this purpose. Another factor which may have influenced these results is the low sampling rate which excluded high frequencies above 8kHz from influencing the performance. MFCCs on the other hand were by far the strongest predictor of sound source elevation despite the low sampling rate used in our stimuli. While the Mean Square Error (MSEs) for predictions using MFCCs were low, they were improved upon when using a combination of MFCCs and CCF. The accuracy in prediction decreased consistently along with increased reverberation time in the testing stimuli. This is an expected result as reverberation adds ambiguity to the testing stimuli.

5. CONCLUSION

The results of this study show that Mel-Frequency Cepstral Coefficients and the Cross-Correlation Function between a pair of binaural signals are good indicators of sound source elevation. Compared to the individual performance of these features, a combination of the two improves the overall performance of an ML algorithm in predicting elevation . Interaural cues used to predict azimuth location do not generally perform well at this task. The algorithm designed in this work could be implemented in the testing of spatial audio presented over headphones as well as in Augmented Reality (AR) and Virtual Reality (VR) applications. There are several ways in which this study could be expanded upon. For instance, additional training stimuli such as music and noise could be used to further evaluate the cues used in this study. In addition to this, implementing a virtual means of performing head-rotation could eliminate front-back confusion and allow for reliable testing across a fully spherical array of source locations. It is hoped that this algorithm or one similar to it could be used to predict the source elevation of real world sounds occurring in real-time.

6. REFERENCES

[1] K. Youssef, S. Argentieri, and J.-L. Zarader, “Towards a systematic study of binaural cues,” in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pp. 1004–1009, IEEE, 2012.

[2] K. Youssef, S. Argentieri, and J.-L. Zarader, “A binaural sound source localization method using auditive cues and vision,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 217–220, IEEE, 2012.

[3] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, “Tori of confusion: Binaural localization cues for sources within reach of a listener,” *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1627–1636, 2000.

[4] B. Kapralos, M. Jenkin, and E. Milios, “Virtual audio systems,” *Presence: Teleoperators and Virtual Environments*, vol. 17, no. 6, pp. 527–549, 2008.

[5] N. Ma, T. May, and G. J. Brown, “Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2444–2453, 2017.

[6] V. R. Algazi, C. Avendano, and R. O. Duda, “Low-frequency ild elevation cues,” *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 2237–2237, 1999.

[7] K. Iida, Y. Ishii, and S. Nishioka, “Personalization of head-related transfer functions in the median plane based on the anthropometry of the listener’s pinnae,” *The Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 317–333, 2014.

[8] C. Lim and R. Duda, “Estimating the azimuth and elevation of a sound source from the output of a cochlear model,” in *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, IEEE Comput. Soc. Press, 1994.

[9] K. D. Martin, "Estimating azimuth and elevation from interaural differences," in *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*, pp. 96–99, Oct 1995.

[10] H. Meutzner, A. Schlesinger, S. Zeiler, and D. Kolossa, "Binaural signal processing for enhanced speech recognition robustness in complex listening environments," *Proc. CHiME-2013, Vancouver, Canada*, pp. 7–12, 2013.

[11] T. Ashby, R. Mason, and T. Brookes, "Prediction of perceived elevation using multiple pseudo-binaural microphones," in *Audio Engineering Society Convention 130*, Audio Engineering Society, 2011.

[12] M. Lovedee-Turner and D. Murphy, "Application of machine learning for the spatial analysis of binaural room impulse responses," *Applied Sciences*, vol. 8, no. 1, 2018.

[13] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[14] G. Kearney and T. Doyle, "An hrtf database for virtual loudspeaker rendering," in *Audio Engineering Society Convention 139*, Audio Engineering Society, 2015.

[15] F. Chollet *et al.*, "Keras," 2015.

[16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning.," in *OSDI*, vol. 16, pp. 265–283, 2016.

[17] D. T. Murphy and S. Shelley, "Openair: An interactive auralization web resource and database," in *Audio Engineering Society Convention 129*, Audio Engineering Society, 2010.

[18] J. Blauert, *Spatial Hearing - Revised Edition: The Psychophysics of Human Sound Localization*. The MIT Press, 1996.

[19] H. O'Dwyer, E. Bates, and F. M. Boland, "A machine learning approach to detecting sound-source elevation in adverse environments," in *Audio Engineering Society Convention 144*, Audio Engineering Society, 2018.