ICA 2022
24th INTERNATIONAL CONGRESS ON ACOUSTICS
October 24(Mon) - 28(Fri), 2022 Gyeongju, Korea

PROCEEDINGS of the
24th International Congress on Acoustics

October 24 to 28, 2022 in Gyeongju, Korea

Predicting room acoustical parameters from running signals using a precedence effect model and deep neural networks

Jeramey Tyler[(1)], Mei Si[(2)], Jonas Braasch[(3)]

[(1)]Rensselaer Polytechnic Institute, Troy, United States, tylerj2@rpi.edu

[(2)]Rensselaer Polytechnic Institute, Troy, United States, sim@rpi.edu

[(3)]Rensselaer Polytechnic Institute, Troy, United States, braasj@rpi.edu

ABSTRACT

A model architecture is presented to predict room acoustical parameters from a running binaural signal. For this purpose, a deep neural network architecture is combined with a precedence effect model to extract the spatial and temporal locations of the direct signal and early reflections. The precedence effect model builds on a modified BICAM algorithm [1], for which the first layer auto-/cross correlation functions are replaced with a Cepstrum method. The latter allows a better separation of features relating to the source signal's early reflections and harmonic structure. The precedence effect model is used to create binaural activity maps that are analyzed by the neural network for pattern recognition. Anechoic orchestral recordings were reverberated by adding four early reflections and late reverberation to test the model. Head-related transfer functions were used to spatialize direct sound and early reflections. The model can identify the main reflection characteristics of a room, offering applications in numerous fields, including room acoustical assessment, acoustical analysis for virtual-reality applications, and modeling of human perception.

Keywords: Precedence effect, auditory modelling, deep learning

1 INTRODUCTION

It is investigated how a binaural model can be used to extract room acoustical features from a running binaural signal consisting of an anechoic orchestral recording with simulated early reflections using a head-related transfer function catalog and late reverberation. At the core of the human ability to extract information from sound sources in reverberant spaces are auditory mechanisms related to the *precedence effect*. The precedence effect, formerly also called the *law of the first wave front*, describes the ability of the auditory system to suppress information about secondary sound sources that are reflected off walls and other objects. This enables the auditory system to localize the actual position of a sound source by making the localization cues pertinent to the direct signal component available.

This is a non-trivial task for the auditory system since the direct signal and the reflected signal parts overlap in time and frequency. The primary cues to localize a sound source are *Interaural Time Differences* (ITDs) and *Interaural Level Differences* (ILDs). ITDs occur because the path lengths between a sound source and both ears differ depending on the incoming azimuth angle. The cross-correlation algorithm is a good algorithm to simulate the processes of the auditory system to extract ITD cues. Unfortunately, the traditional cross-correlation model fails in reverberant conditions. Recently, Braasch proposed a binaural model based on a second-layer cross-correlation algorithm [1].

This paper presents an adapted method using the cepstrum algorithm to estimate the location and degree of early reflections. The cepstrum method has been used before to dereverberate signals [5]. For atonal signals, the resulting model performs as well as the previously published BICAM model [1], but its performance with tonal signals is much improved. The cepstrum method is an example of homomorphic filters that use non-linear processes to separate features that occupy the same feature space. In this project, we were able to use

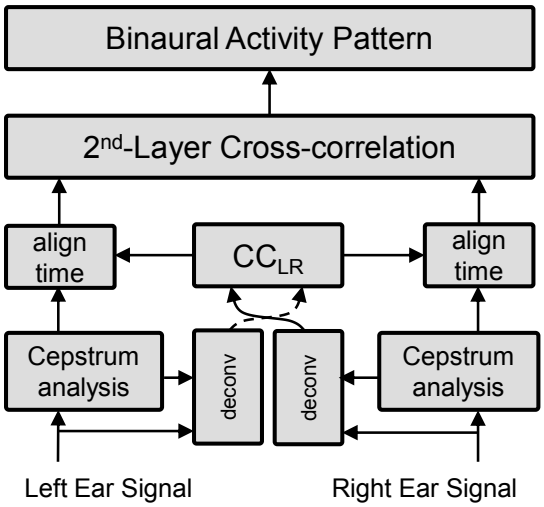

Figure 1. Architecture of the cepstrum-based binaural model.

a cepstrum-based method to separate the features from specular reflections from those induced by the signals' periodicities. A deep learning model is then used to analyze the output of the binaural model. Deep learning methods have been used successfully in the past for sound localization with binaural models [6, 9, 2].

2 CEPSTRUM METHOD

The cepstrum algorithm is a standard method in speech technology to separate the carrier signal $S(f)$ from the speech envelope $H(f)$, for example, to extract speech formants. The frequency spectrum of a speech signal is defined as:

$$G(F) = S(f) \cdot H(f). \tag{1}$$

By taking the logarithm of the speech spectrum, the multiplication of $S(f)$ and $H(f)$ becomes an addition:

$$\log|G(f)| = \log|S(f)| + \log|H(f)|. \tag{2}$$

After performing a Fourier transform of the logarithmized speech spectrum, the cepstrum is obtained:

$$C(\tau) = \mathscr{F}^{-1}\{\log(G(f))\} = \mathscr{F}^{-1}\{\log(\mathscr{F}\{g(t)\})\}. \tag{3}$$

The variable $n\tau$ is the so-called *quefrency*, which has a unit in seconds while not being identical to time due to the nature of the applied logarithm. For voiced signals, we will find a peak at $\tau_0 = 1/f_0$ for the fundamental frequency of the source signal. This part is subtracted through linear filtering. The formants are then visualized by applying the inverse Fourier transformation.

The algorithm resembles a dual-layer, spatio-temporal filter to separate auditory features for the direct and reverberant signal parts of a running signal. Figure 1 shows the architecture of the model, with a signal flow ascending from bottom to top. In the first stage, the model performs a cepstrum analysis for the left and right ear signals. For each channel, the result of the cepstrum analysis is used to compute a deconvolution filter, which is then used to deconvolve the ear signal (see boxes "deconv"). Next, the deconvolved signals are cross-correlated to determine the interaural time difference (ITD) of the direct signal (box: "CC_{LR}"). The estimated ITD is used to time align the estimated impulse responses from the cepstrum analysis (boxes: "align time"). These two time-aligned impulse responses are cross-correlated a second time to compute a binaural activity pattern. The generated patterns are fed to the deep learning model.

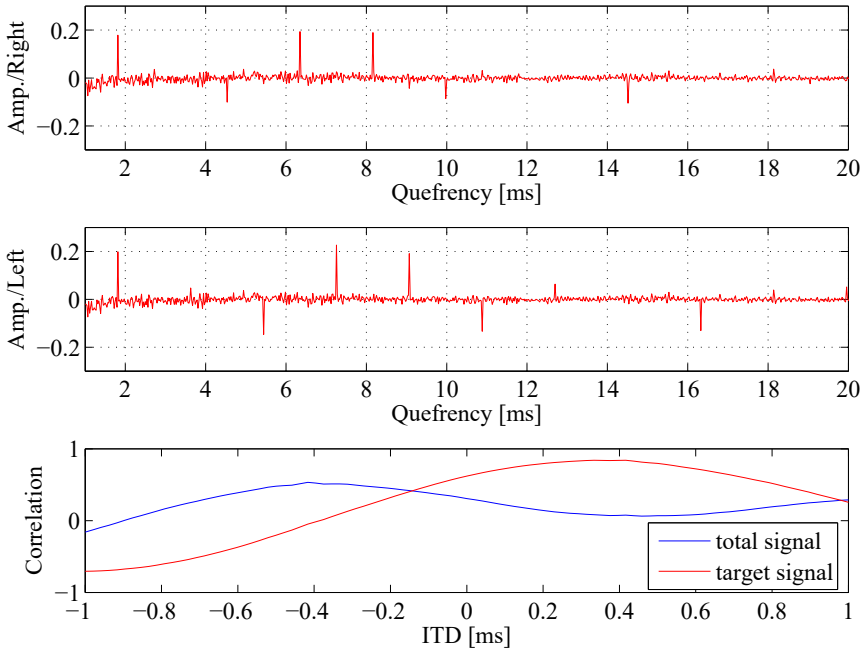

Figure 2. Cepstrum-based localization model example to calculate the interaural time difference (ITD).

Figure 2 shows a practical example of the binaural model performance to determine the interaural time difference (ITD) for the direct signal, by estimating deconvolution filters from the cepstra of the left and right signals of a binaural signal. This approach is depicted for a speech signal with a direct signal (0.4-ms ITD) and two reflections (−0.4-ms ITD each, with delays of 7 and 9 ms with respect to the direct signal, each reflection has the same amplitude as the direct signal). The top graph shows the cepstrum for the left signal, and the center graph illustrates the cepstrum for the right signal. The cepstra are used as deconvolution filters for the left and right signals before the cross-correlation function is obtained. The bottom graph shows that this method allows us to calculate the lateral position of the target sound correctly (red curve). The blue curve shows the cross-correlation function for the total signal, which has its maximum near −0.4-ms because the combined amplitude of the reflections is twice as high as the amplitude of the direct sound.

We can now apply the cepstrum method to extract/visualize the locations of early reflections. The cepstrum allows us to separate the spectral harmonics of the signal (periodicity) and the reflections, which is possible if the early reflections do not come in too early – see Fig. 3.

3 RAY TRACING

Generic room impulse responses were generated using the approach above to test the general suitability of the binaural model. Each impulse response contained four individual early reflections and late reverberation. The late reverberation tail is formed by a stochastic process with an underlying Gaussian distribution with an exponentially-decaying window adjusted to reverberation times (RT). The parameters for each test condition are given in the result section. The HRTFs were taken from [3]. The impulse responses were auralized using anechoic orchestral recordings [4]. The complete set of recordings was segmented into 20-s long samples resulting in 76 samples total. The samples were convolved with the binaural room impulse responses and analyzed by the binaural model.

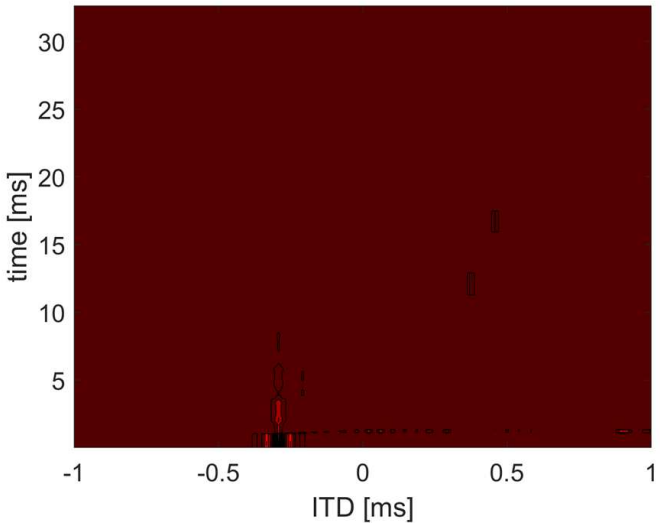

Figure 3. Example Binaural Activity Map for the cepstrum-based binaural model using an anechoic Mozart Clip, with direct signal (Amplitude: 1, ITD: −0.3 ms and two early reflections – Reflection 1: 0.8 Amplitude, 13-ms ISI, 0.4-ms ITD, Reflection 2: 0.8 Amplitude, 18-ms ISI, 0.5-ms ITD.

4 RESULTS

Figure 4 shows four examples of binaural activity maps calculated from the running signals analyzed by the binaural model. The spatial location of the first reflection was varied in each of the figures. The remaining parameters were as follows: direct signal: 0° azimuth, other Early Reflection (ER) azimuths: 315°, 60°, 300°; ER delays: 20 ms 26 ms, 36 ms, 44 ms; other ER amplitudes: 0.8, 0.6, 0.55, 0.55; RT=2 s, RT amp.=0.25.

Next, the model temporal location of the first reflection and its relative amplitude was investigated. All data points show the mean and standard deviations over all 76 samples. The left graph of Fig. 5 shows the results for 5 different initial time delay gaps (ITDG), demonstrating that the model is able to estimate the ITDG accurately (direct signal: 0° azimuth, Early Reflection (ER) azimuths: 315°, 300°, 60°, 300°; other ER delays: 27 ms 30 ms; ER amplitudes: 0.8, 0.75, 0.15, 0.14; RT=2, s RT amp.=0.25). Estimating the reflection amplitude was more difficult. A monotonic relationship can be found between the actual and the estimated amplitude; the errors are fairly high. It should be noted that the ITDG measurement is part of the ISO 3382 metric catalog, while the first reflection amplitude is not.

5 DEEP LEARNING

The goal of the neural network analysis is to automatically detect the direction of the direct sound and the spatial/temporal locations of early reflections and the strength of these reflections. This work builds on an earlier neural model that used the BICAM algorithm [2]. The model was able to detect a reflection in the presence of a direct sound with the following accuracies: time-delay of reflection (98.5%–99.6%), direct source lateralization (93.0%–94.9%), and reflection lateralization (84.1%–86.1%). The idea for the current study is that the output of the cepstrum-based precedence effect is robust enough to extract the parameters from a complete binaural activity map to read the underlying binaural activity patterns.

The new convolutional neural network was implemented based on the VGG architecture [8]. The neural network is fed binaural activity maps created by the binaural cepstrum model in the form of images with 256×256 pixels resolution. A large synthetic dataset was generated to train the neural network. In a pilot study, a dataset of 10,000 audio samples was created with a direct sound source and a computer-generated single reflection. Each anechoic sample had a duration of 10 seconds at 24-bit resolution and 48-kHz sampling frequency. Direct signals and reflections were spatialized using head-related transfer functions (HRTFs) using

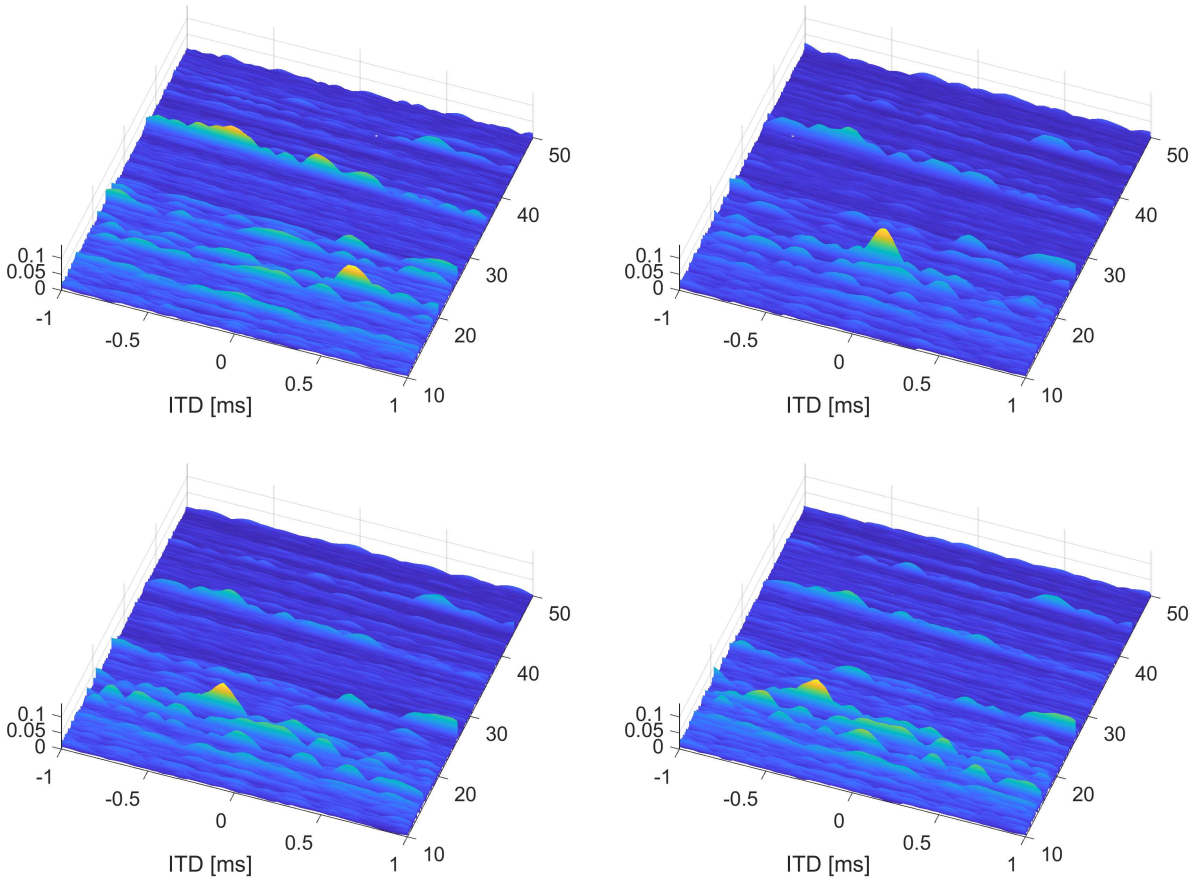

Figure 4. Binaural activity maps calculated from a running signal. For each panel, the azimuth of the first reflection was varied: 60° (top-left panel), 0° (top-right panel), −45° (bottom-left panel), −60° (bottom-right panel). All other parameters were kept constant. For more details see text.

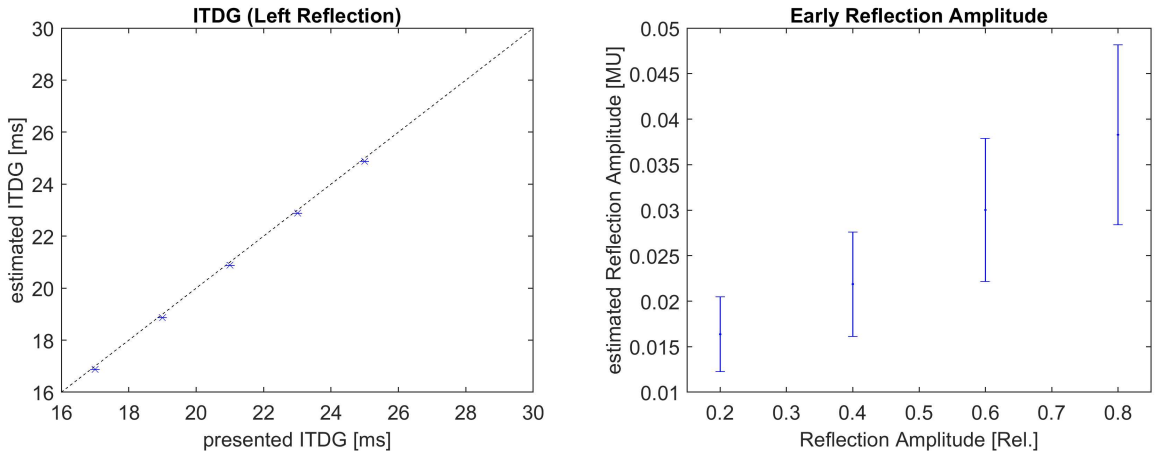

Figure 5. Left: Calculated initial time delay gap; Right: Calculated reflection amplitude.

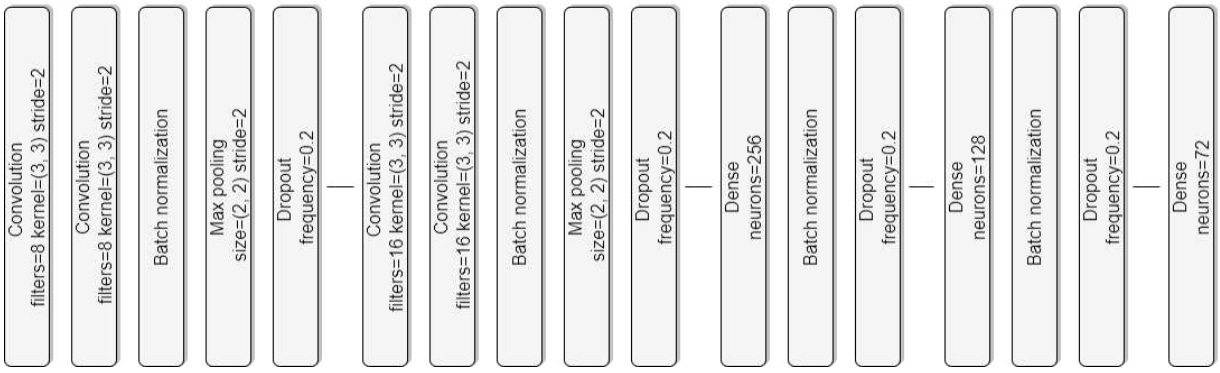

Figure 6. Neural network architecture.

Kemar dummy-head data [3]. Both azimuths for direct sound and reflection and the reflection amplitude and delay were randomized.

A dataset of synthetic reflections was generated using HRTF measurements [3] and anechoic orchestral recordings [7]. Samples were generated using 72 HRTF measurements taken at a 0° elevation with azimuths in the range 0–355° in 5° intervals. The orchestral recordings were randomly sampled for ten-second segments, which were then convolved with a randomly chosen HRTF measurement in order to change the perceived direction of the direct sound. A single synthetic reflection was added at a randomly chosen HRTF with an amplitude between 0.5 and 0.8 and a delay between 1 and 10 ms. Room reverberation was added with a delay between 1 and 10 ms, a reverberation time between 1 and 10 ms, and an amplitude between 0.05 and 0.08. A total of 10,000 samples were generated and randomly separated into training, testing, and validation sets with a 60%/20%/20% split.

The convolutional neural network employed an architecture based on the VGG architecture [8] with two convolution groups followed by two fully connected groups, see Fig. 6. Convolution groups consisted of two convolution layers with a kernel size of 3×3 and a stride of 2, batch normalization, max pooling with a pooling size of 2×2 and a stride of 2, and a dropout layer with a frequency of 0.2 and employed a rectified linear unit activation function. Convolution layers in the first group used 8 filters and those in the second group used 16 filters. The fully connected groups were composed of a fully connected layer, batch normalization, and a dropout layer with a frequency of 0.2 and employed a rectified linear unit activation function. The fully connected layer in the first group had 256 neurons, whereas the second fully connected layer had 128 neurons. The output layer had 72 neurons corresponding to the 72 possible azimuths and employed a categorical cross-entropy activation function.

The neural network was used to make three predictions about the azimuth of the direct sound: whether the azimuth was to the left or the right, whether the azimuth was in front or behind, and the degree of the azimuth. The same neural network was used for each prediction with minor changes to the output layer. For predicting the azimuth, the output layer was composed of 72 neurons and used a categorical cross-entropy activation function, while predicting left or right, and front or back were both composed of a single neuron and used a binary cross-entropy activation function. At the task of predicting left or right the network was trained for 10 epochs and achieved a peak training accuracy of 94%, a peak validation accuracy of 99%, and a peak test accuracy of 98% where left was defined as azimuth ≥ 180 and right as azimuth < 180. At predicting front or back, the network was again trained for 10 epochs and achieved a peak training accuracy of 74%, a peak validation accuracy of 75%, and a peak test accuracy of 76% where front was defined as azimuth < 90 or azimuth ≥ 270 and back was defined as $90 \leq$ azimuth < 270. Finally, for the task of predicting azimuth, the network was allowed to train for 50 epochs and achieved a peak training accuracy of 55%, a validation accuracy of 57%, and a test accuracy of 59%. It is our belief that the network's difficulty predicting front or back limited the accuracy of predicting azimuth. In order to test this hypothesis, we intend to generate the data again with only azimuths occurring in the front.

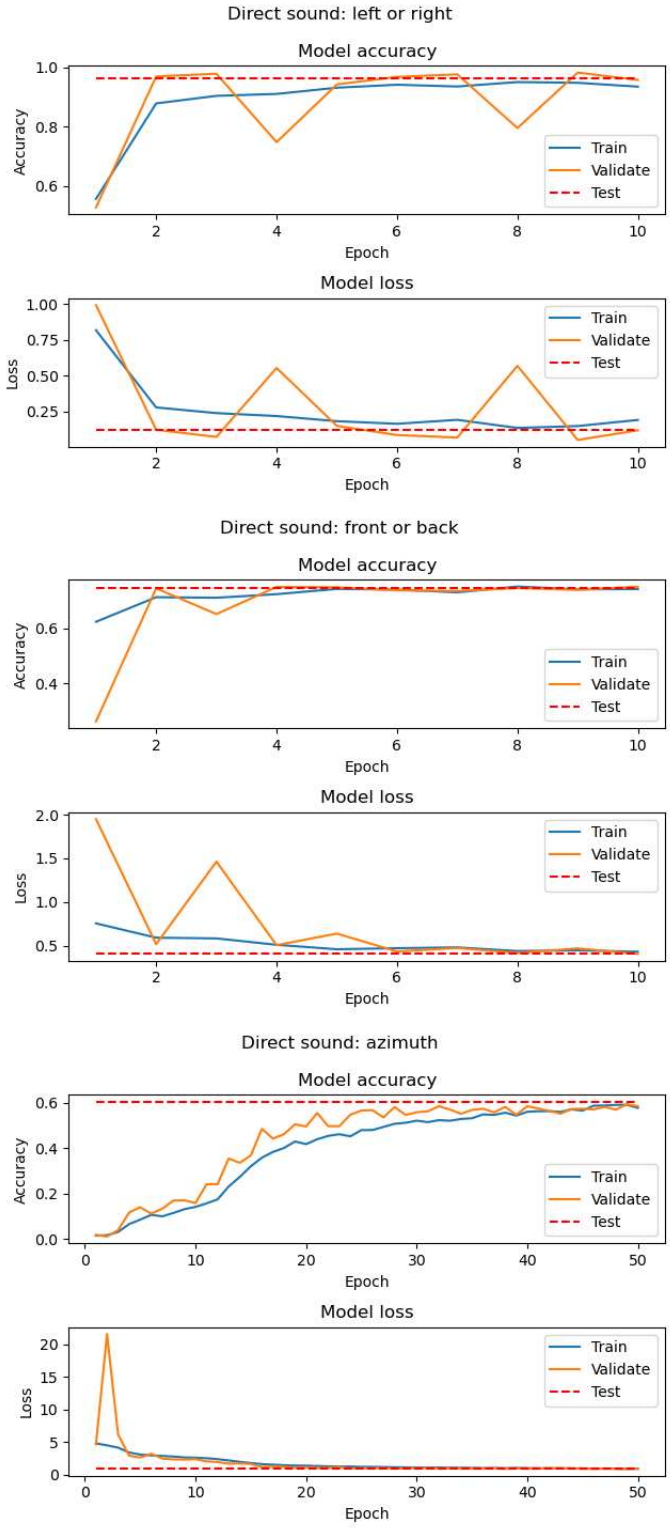

Figure 7. Results from direct sound predictions. Our network achieved 98% accuracy in predicting if the sound occurred to the left or right, 80% predicting if the sound occurred in front of or behind, and 59% predicting the azimuth of the sound.

6 CONCLUSIONS

A new precedence effect model is proposed to extend the capabilities and robustness of a previous model by replacing the the first cross-correlation stage of a dual-layer cross-correlation algorithm with a cepstrum-based approach. The model is promising and appears to be less vulnerable to periodic signal components. A deep learning model has been added on top of the binaural model and early results are promising. The investigation will be continued to improve the performance of the deep learning model.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. IIS-1909229.

REFERENCES

[1] J. Braasch. Binaurally integrated cross-correlation/auto-correlation mechanism (bicam). *The Journal of the Acoustical Society of America*, 139(4):2211–2211, 2016.

[2] N. Deshpande and J. Braasch. Detection of early reflections from a binaural activity map using neural networks. *The Journal of the Acoustical Society of America*, 146(4):2529–2539, 2019.

[3] W. G. Gardner and K. D. Martin. HRTF measurements of a KEMAR. *The Journal of the Acoustical Society of America*, 97(6):3907–3908, 1995.

[4] T. Hidaka. Recording of anechoic orchestral music and measurement of its physical characteristics based on the auto-correlation function. *Acustica*, 67:68–70, 1988.

[5] K. Kumar and R. M. Stern. Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4282–4285. IEEE, 2010.

[6] 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*, 25(12):2444–2453, 2017.

[7] J. Pätynen, V. Pulkki, and T. Lokki. Anechoic recording system for symphony orchestra. *Acta Acustica united with Acustica*, 94(6):856–865, 2008.

[8] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[9] Z.-Q. Wang, X. Zhang, and D. Wang. Robust speaker localization guided by deep learning-based time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(1):178–188, 2018.