Predicting Ensemble Width and Location in Binaural Recordings of Music with Convolutional Neural Networks

Paweł Antoniuk*[1] and Sławomir Zieliński[1]

[1]Faculty of Computer Science, Białystok University of Technology, 15-351 Białystok, Poland; pawel.antoniuk@sd.pb.edu.pl
*Corresponding author: pawel.antoniuk@sd.pb.edu.pl

Abstract

Binaural audio technology has been around for many years, but its popularity has dramatically increased over the past decade due to advancements in virtual reality and streaming technologies. Along with its growing popularity, the quantity of publicly accessible binaural audio materials has also expanded. Consequently, there is now a need for automated and objective measurements of spatial content information, with ensemble width and location being the most important. In this study, 30 head-related transfer functions and 192 binaural music recordings from publicly accessible multi-track recording repositories were used to synthesize 23,040 binaural recordings. The synthesised recordings were then used to train a convolutional neural network prediction model, the aim of which was to predict the width and location of the ensemble for unseen recordings. The results indicate that models of ensemble breadth and position can be successfully constructed with low prediction errors — 4.62° (±0.09°) degrees for ensemble location and 8.63° (±0.29°) degrees for ensemble width. This approach is the first of its kind to predict both ensemble width and location, offering a more accurate representation of spatial properties. This suggests significant potential for advancing spatial audio applications in virtual reality and streaming technologies, by providing audio engineers with tools that can leverage these methods to enhance spatial audio experiences.

1 Introduction

2 Related studies

3 Methodology

Experiments in this study were conducted on 2340 binaural recordings of music. The binaural recordings were synthesized semi-automatically using 192 multi-track publicly-available music recordings and 30 HRTF databases. For each multi-track recording and HRTF database pair, four binaural recordings were synthesized for different random ensemble parameters — its location ϕ and width θ — as defined in Section 3.1. Both parameters were drawn from uniform random distribution.

3.1 Ensemble location and width definition

The primary objective of the model developed in this study is to predict the ensemble location (θ) and width (ω), as illustrated in Figure 1. The ensemble is defined as a group of audio point sources. The ensemble width ω is defined as the angular width between two extreme point sources, while the ensemble location θ represents the middle angle between two extreme sound sources.

3.2 Synthesis of binaural music recordings

3.3 Spectrogram conversions of binaural music recordings

Prior to being fed into the model, the binaural recordings of music were transformed into magnitude spectrograms. Figure 2 illustrates example spectrograms for four distinct binaural recordings of music. It should be noted that the original spectrograms were stored in a floating-point-precision matrix format.

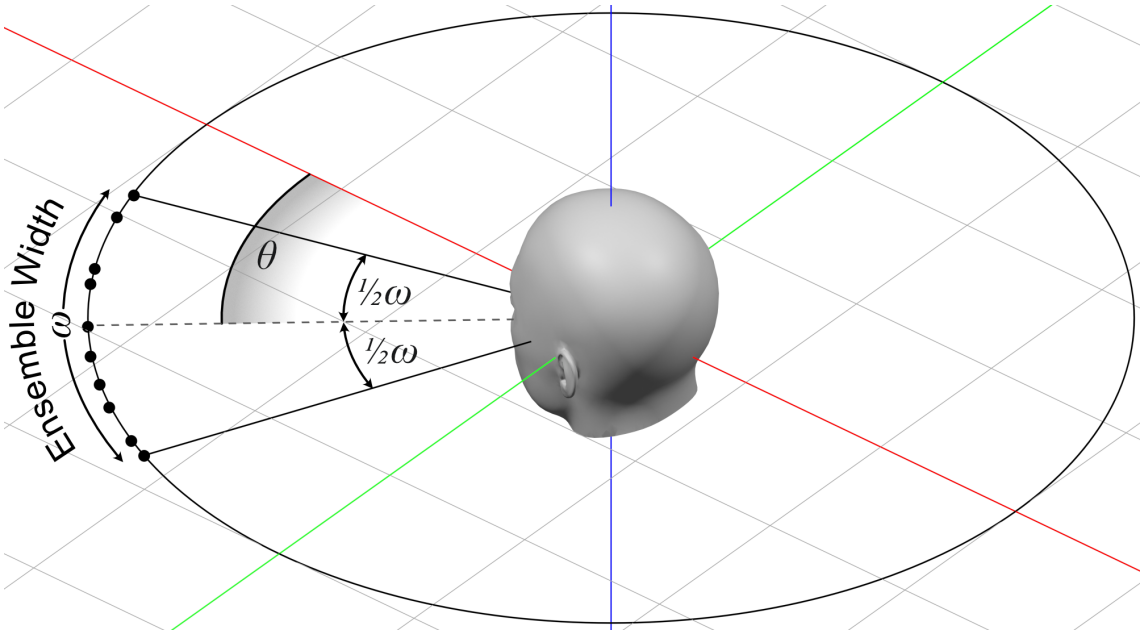

Figure 1: Illustration of ensemble width (ω) and ensemble location (θ) relative to the direction of the head. Black dots represent the positions of audio sources. The ensemble location (θ) is defined as the angular position of the center of the ensemble relative to the direction the head is turned. The ensemble width (ω) is defined as angular distance between two extreme audio sources.

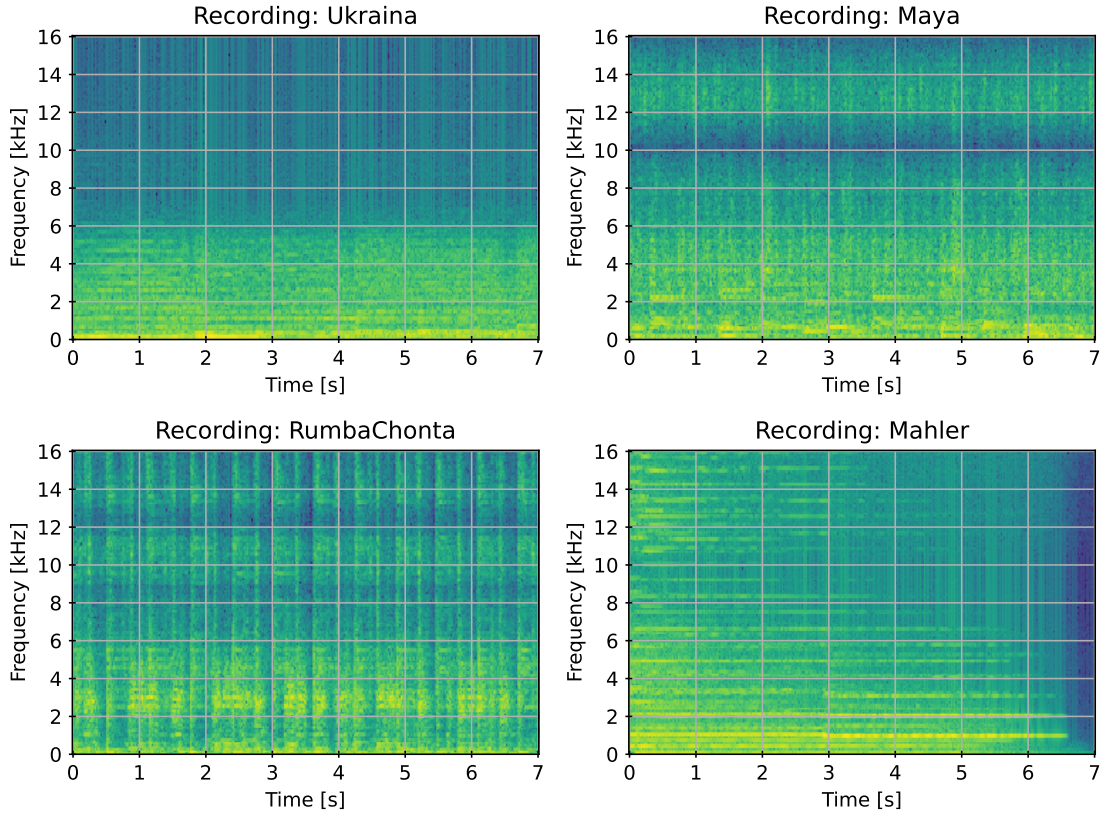

Figure 2: An illustrative example of the magnitude spectrograms generated from synthesized binaural recordings of music. Each spectrogram was derived from a different randomly selected binaural music recording. These spectrograms were subsequently employed as input data for training and assessing the convolutional neural network (CNN) in this study.

3.4 Convolutional neural network topology

3.5 Model training and evaluation

4 Results

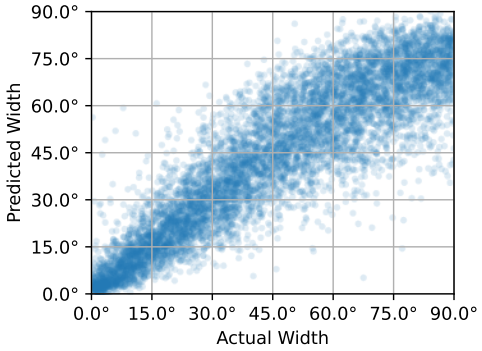

Figure 3: A comparison between the actual and the predicted ensemble width for a single iteration (of the total five)

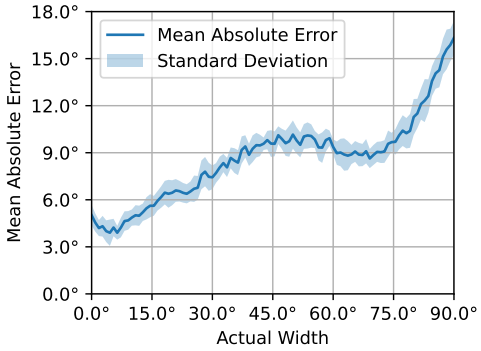

Figure 4: The impact of the actual ensemble width on the mean absolute prediction error, averaged across all five iterations, with indicated standard deviation.

Figure 3 illustrates the comparison between the actual and the predicted ensemble width. The results demonstrate that the model exhibits optimal prediction quality for narrower ensemble widths (X° for Y°) and that its performance deteriorates with the increase of the ensemble width (Z° for W°). Figure 4 further demonstrates that the relationship between the prediction error and the actual width is not linear, exhibiting a depression between 60 and 75 degrees. This suggests, that the ensemble width has a significant impact on the ensemble width estimation.

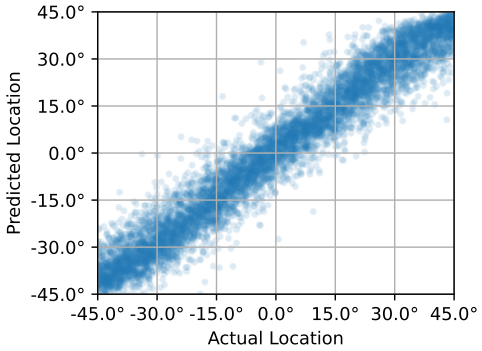

Figure 5: A comparison between the actual and the predicted ensemble location for a single iteration (of the total five)

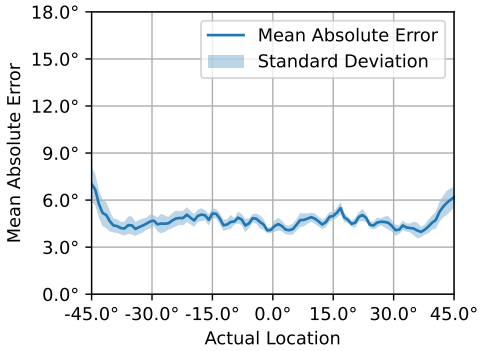

Figure 6: The impact of the actual ensemble width on the mean absolute prediction error, averaged across all five iterations, with indicated standard deviation.

In contrast to the correlation between ensemble width and its prediction error, there is no significant relationship between the actual location and its prediction error, as illustrated in Figures 5 and 6. This finding indicates that the model's capabilities for localizing the center of the ensemble is robust, unaffected by the actual spatial positioning of the ensemble, including lateral locations.

Figure 7 shows influence of both the ensemble width and location on mean absolute error for ensemble location. This shows more detailed view of data previously presented on Figure 6. The figure shows that there is a slight asymmetry.

Similarly, Figure 8 shows influence of both the ensemble width and location on mean absolute error for ensemble width.

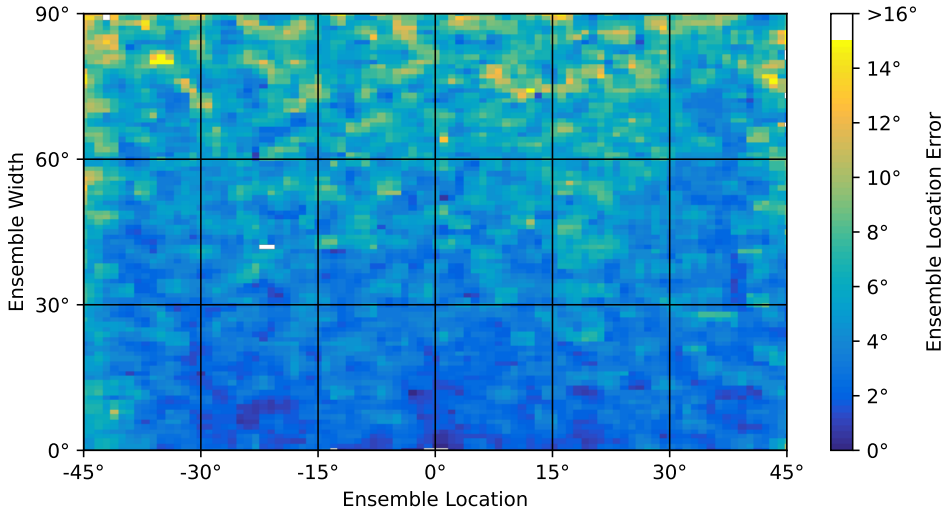

Figure 7: The heatmap that illustrates the mean absolute error (MAE) of ensemble location distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors.

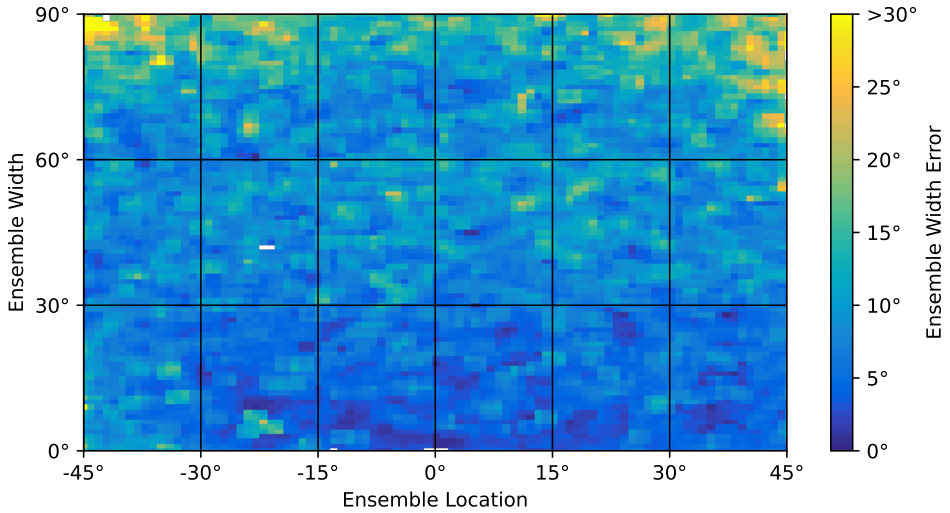

Figure 8: The heatmap that illustrates the mean absolute error (MAE) of ensemble width distribution across different ensemble locations (x-axis) and ensemble widths (y-axis). The color intensity corresponds to the MAE values, with lighter areas indicating higher errors. Notably, the values between 30° and 60° on the y-axis exhibit unexpectedly higher MAE values region — please see Figure 4 for comparison.

5 Conclusions

6 References

[Gre93]

References

[Gre93] George D. Greenwade. The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3):342–351, 1993.