H. O'Dwyer and F. Boland, "HRTF Clustering
for Robust Training of a DNN for Sound Source Localization"
J. Audio Eng. Soc., vol. 70, no. 12, pp. 1015–1026, (2022 December).
DOI: https://doi.org/10.17743/jaes.2022.0051

HRTF Clustering for Robust Training of a DNN for Sound Source Localization

HUGH O'DWYER,* *AES Student Member,* **AND FRANCIS BOLAND,** *AES Member*
(odwyerh@tcd.ie) (fboland@tcd.ie)

Trinity College Dublin, Dublin, Ireland

This study shows how spherical sound source localization of binaural audio signals in the mismatched head-related transfer function (HRTF) condition can be improved by implementing HRTF clustering when using machine learning. A new feature set of cross-correlation function, interaural level difference, and Gammatone cepstral coefficients is introduced and shown to outperform state-of-the-art methods in vertical localization in the mismatched HRTF condition by up to 5%. By examining the performance of Deep Neural Networks trained on single HRTF sets from the CIPIC database on other HRTFs, it is shown that HRTF sets can be clustered into groups of similar HRTFs. This results in the formulation of central HRTF sets representative of their specific cluster. By training a machine learning algorithm on these central HRTFs, it is shown that a more robust algorithm can be trained capable of improving sound source localization accuracy by up to 13% in the mismatched HRTF condition. Concurrently, localization accuracy is decreased by approximately 6% in the matched HRTF condition, which accounts for less than 9% of all test conditions. Results demonstrate that HRTF clustering can vastly improve the robustness of binaural sound source localization to unseen HRTF conditions.

0 INTRODUCTION

Binaural sound source localization (SSL) is a complex task that involves extracting source direction information from the signals arriving at each ear. Rayleigh in his 1907 *Duplex Theory* [1] proposed that the ability to localize the direction to a sound source is based on perception of the interaural time difference (ITD) and the interaural sound level difference (ILD) between sounds at each ear. Understanding how the ITD and ILD cues are combined and how conflicting cues are traded has motivated continuing research in the duplex theory; see, for example, the 2016 transaural experimental work of Hartmann et al. [2] and 2013 neuroscience research of Edmunds et al. [3] that explores whether ITD and ILD are represented by independent or integrated codes in the human auditory cortex. The duplex theory supports the understanding of how these cues inform perception of a source's location on the horizontal plane.

Perception of a source's location along the vertical plane is understood to rely especially on spectral cues resulting from the reflection of sound waves off an individual's body—particularly the torso, shoulders, and pinna folds [4]. A phenomenon known as the pitch-height effect describes the apparent rise in volume of high-frequency components of a sound source as its position relative to a listener increases [5]. A 2005 study by Raykar showed that as a sound source increased in height relative to a listener, so too do resonant peak frequencies, [6] which inform understanding of the height of a sound source.

Although this process has been studied widely, it is not fully understood how binaural cues are mapped to source location by the mammalian brain, particularly in relation to the vertical plane. It is known, however, that the cues that allow an individual to perform localization of the direction in azimuth and elevation of a sound source are contained within their Head-Related Transfer Function (HRTF). The HRTFs' spectrograms allow for a visualization of the spectral peaks and notches at frequencies that vary with source elevation. These filter effects have been a well-studied area since the 1990s [7, 8]. They were described in detail by Cheng and Wakefield in the *Journal of the Audio Engineering Society* [9], in which their role in the synthesis of spatial sound over headphones was explored. However, it is not known how densely the azimuth and elevation space must be sampled, in terms of locations, to generate an HRTF set that sufficiently captures and represents an individual's listening.

A SSL algorithm that uses just two sensors, in the same way humans do, would be of interest in the area of machine perception. For example, it could be used for acoustic event

*To whom correspondence should be addressed e-mail: odwyerh@tcd.ie

detection and localization. Such an algorithm could also be used to assist hard of hearing people or used within an augmented reality system [10, 11]. Other uses may include humanoid robotic systems and target tracking systems [12–14].

To develop such an algorithm with Machine Learning (ML) and specifically a Deep Neural Network (DNN), there are a number of challenges that must be addressed. These include how to create a large enough database of binaural audio signals from different source positions for training and testing purposes. Then there is the issue of how to decide what, if any, features should be extracted from the binaural audio to provide cues to the DNN of the source direction. The most effective method of extraction and representation of these features must also be determined. A database of binaural recordings must be created consisting of recordings synthesized from multiple HRTF measurements. Two testing conditions must then be considered; the first is the *matched condition*, in which the algorithm is both trained and tested on recordings generated from the same HRTF set, and the second is the *mismatched condition*, in which training is performed with one HRTF set and tested on another. The research reported here examines the choice and performance of features in the mismatched condition.

Many studies of ML algorithms for SSL use only one or a small number of individuals to provide recordings for both testing and training. The commonly adopted approach to generate a set of binaural recordings is through synthesis using an individual's HRTF set so the matched condition is equivalent to training and testing using signals synthesized using the same HRTF set, whereas mismatched relates to training and testing on signals synthesized from different HRTF sets. The individuality of HRTFs does mean that for an ML algorithm to be trained to be robust in the mismatched condition, consideration must be taken of the choice of features and the number and nature of the HRTF sets used in the training of the system.

The research reported in this paper explores how cluster analysis, of the performance of a DNN for binaural SSL trained on multiple HRTF sets, may provide an improved performance in the mismatched case. In SEC. 1, a description is given of the creation of a synthesized database of binaural audio files used in this study. In SEC. 2, binaural cues and their extraction from these synthesized signals are explained. SEC. 3 presents ML algorithms that use binaural cues to estimate source location within a binaural signal. A novel feature combination and ML architecture are presented by the authors and compared with two state-of-the-art methods. In SEC. 4 it is shown that the results gathered by the proposed algorithm can be improved upon using HRTF clustering. Here, a method of clustering together similar HRTFs is presented. Training on a combination of central HRTFs from within these clusters is then shown to greatly improve results in the mismatched condition. A discussion of these results is presented in SEC. 5, followed by a conclusion in SEC. 6. Here, it is shown that HRTF clustering can improve vertical localization in binaural signals by up to 13% in the mismatched condition.

1 BINAURAL SIGNAL MODEL AND SYNTHESIS

Binaural signals can be synthesized by convolving a two-channel Head-Related Impulse Response (HRIR) or Binaural Room Impulse Response (BRIR) with a source signal. An HRTF database, for example, the CIPIC database [15], has pairs of left and right HRIRs, for a source located in direction $\mathbf{\Theta} = (\theta, \phi)$ θ and ϕ refer to the azimuth and elevation direction to the source, respectively, referenced to the center of the listener's head. The HRIRs are notated as $h_l(t, \mathbf{\Theta})$ and $h_r(t, \mathbf{\Theta})$ and if there is a source signal *s(t),* then the signals received at the binaural sensors can be represented as

$$\begin{aligned} x_l(t) &= h_l(t, \mathbf{\Theta}) * s(t) + n_l(t) \\ x_r(t) &= h_r(t, \mathbf{\Theta}) * s(t) + n_r(t) \end{aligned}, \tag{1}$$

where $n_{l/r}$ represent additive noise terms and t represents time. The received signals can then be interpreted in the frequency domain as

$$\begin{aligned} X_l(f) &= H_l(f, \mathbf{\Theta})S(f) + N_l(f) \\ X_r(f) &= H_r(f, \mathbf{\Theta})S(f) + N_r(f) \end{aligned}. \tag{2}$$

In this study, HRIRs from the CIPIC database [15] are convolved with speech signals from the telecommunications and signal processing (TSP) database [16]. The CIPIC database consists of sets of HRTF measurements from 45 subjects including two HRTF sets measured on a KEMAR manikin, representative of the average male and female ear sizes, respectively. Each HRTF set in the CIPIC database contains measurements at 25 azimuthal positions ranging from -80° to 80°, i.e., from left to right. At each azimuthal position, there are 50 HRTF measurements at different elevations ranging from -45° to 230.625°. These can be referred to as sagittal planes or the median plane in the instance in which $\theta = 0^\circ$. Each HRTF measurement is represented as a 200-tap impulse response saved as a ".wav" file with a sampling rate of 44.1 kHz.

The TSP laboratory at McGill University created the TSP speech database in 2002 [16]. It consists of 1,400 utterances spoken by 24 speakers, half of which are male. The database was recorded at a sampling rate of 48 kHz in an anechoic room. The TSP database was chosen in this study over the more popular TIMIT speech database [17] because of its higher sampling rate in recording. This should ensure that any high-frequency content in the speech recordings above 8 kHz is maintained. This is important because it has been shown that notch frequencies above 7 kHz are significant cues for vertical localization [18, 19]. Signals from the TSP database are down-sampled to 44.1 kHz prior to convolution with HRIRs from the CIPIC database to ensure a consistent sampling rate.

For each location within the HRIR set, a total of 100 randomly selected speech samples from the TSP database is used to create a new synthesized binaural signal. Because the authors are concerned with only the front-hemifield, this results in 25 horizontal positions and 25 vertical positions for a total of 625 HRIRs and 62,500 synthesized binaural signals per HRIR set.

2 BINAURAL HEARING AND SOURCE DIRECTION CUES

Beyond interpreting the loudness, pitch, and timbre of sound, the human auditory system is also capable of locating a sound source in 3D space. As explained earlier, the ability to localize the direction to a sound source is understood to be dependent on a variety of interaural and monaural cues, which can often be subtle and may sometimes be contradictory. The nature of some cues may include responses associated with individual characteristics such as the geometry of the pinna or head shape. So, when planning a training strategy for an ML system to learn to robustly estimate the direction to a sound source, the authors seek to avoid individualized cues.

In [20], Lyon explains and discusses the notion of *feature engineering*. This is the process of designing representations or formulating models of signal features that provide an efficient dimensionality reduction from the raw binaural waveform into forms that are well suited as inputs to an ML system. Now, to some ML experts, this step is seen as unnecessary or counterproductive because they advocate that this should be part of the learning process. In this study, the authors adopt a feature engineering approach that seeks to identify features, known from the extensive research literature, to be good cues to the source direction; see studies in SSL as summarized in Jens Blauert's essential text *Spatial Hearing - The Psychophysics of Human Sound Localization* [21, 19]. The authors have adopted two well-known interaural directional features of binaural audio based on the *level difference* and *time difference* and monoaural features extracted by an auditory filter model for the left and right ears.

The extraction of these features from recorded binaural signals facilitates the training of a DNN to provide an estimate of the direction to the sound source. If the binaural signal waveforms were to have been adopted as the inputs, then a more complex architecture accommodating convolutional and spectral processing would be required.

2.1 ILD Feature

ILD is determined by the angle of incidence of a sound wave and the shadowing effect of the head. It is a measurement of the difference in signal level between the left and right ears of a binaural signal and can be measured as a single value over the entire frequency range or as an array of values measured at different frequencies. ILD is a result of high-frequency attenuation caused by the head as sound waves travel around it. For this reason, ILD values below 1.5 kHz are relatively minimal [21].

In this study, ILD is calculated by transforming both left and right signals into the frequency domain. The resultant magnitude values are defined as X_l and X_r and ILD is calculated as follows:

$$ILD(\boldsymbol{\Theta}, f) = 20\log_{10}|\frac{X_l(\boldsymbol{\Theta}, f)}{X_r(\boldsymbol{\Theta}, f)}|. \quad (3)$$

In this study, a single value ILD is used taken as the average value of $ILD(\boldsymbol{\Theta}, f)$ in the frequency range between 1.5 and 20 kHz.

2.2 Cross-Correlation Function

ITD is the most commonly used time-dependent cue in studies on SSL and is a measure of the difference in the time of arrival of a sound at both ears. Although ITD is a reliable localization cue, it has been shown by Ma et al. in [22] that the normalized cross-correlation function (CCF) is a superior cue for training ML algorithms as it contains more information related to the position of a sound source.

In this study, features are extracted by first dividing both left and right signals into frames of length 50 ms with a 25-ms overlap using a Hamming window. The normalized CCF for each frame is then calculated as follows:

$$CCF(k, \tau) = \frac{\sum_m (x_{l,k}(m) - \bar{x}_{l,k})(x_{r,k}(m-\tau) - \bar{x}_{r,k})}{\sqrt{\sum_m (x_{l,k}(m) - \bar{x}_{l,k})^2}\sqrt{\sum_m (x_{r,k}(m-\tau) - \bar{x}_{r,k})^2}}, \quad (4)$$

where $x_{l,k}$ and $x_{r,k}$ refer to the left and right signals, k refers to the time frame index, m refers to the sample index, τ refers to the sample lag, and $\bar{x}_{l,k}$ and $\bar{x}_{r,k}$ refer to the mean values in one frame. Considering the radius of the human head and the speed of sound, the range for the *interaural time delay* is [–1,1] ms. Because the sampling rate used in this study is 44.1 kHz, the range of the sample lag is [–45,45]; resulting in a dimensionality of the CCF feature being 91.

2.3 Gammatone Cepstral Coefficients

Gammatone cepstral coefficients are common features used in recognition tasks and computational auditory scene analysis. This study investigated the use of Gammatone Filter Cepstral Coefficients (GTCCs) as features providing source direction cues to an ML system. The Gammatone filter-bank models the cochlea, which may be assumed to isolate some directional sensitive filtering effects of the outer ear, head, and torso. The cepstral analysis models the loudness perception and decorrelates the per-channel signals. Although the performance of GTCCs has proven to be invaluable cues in tasks such as speech recognition [23] and audio classification [24], their effectiveness in SSL tasks has yet to be fully determined.

Although there are several variations in the methods used to calculate GTCCs from an audio signal, this study will use the methods implemented in MATLAB's "gtcc" function. GTCCs are extracted when the Filter Domain is set to "Time" in MATLAB using the following steps:

- First, the two binaural signals are passed through a 64-channel Gammatone filter-bank.
- For each channel in the filter-bank, the response is then fully rectified, i.e., its absolute value is taken and the signal is split into n-frames according to window size and overlap.
- This creates a time-frequency representation that is a variant of a cochleagram.

- Nonlinear rectification is then calculated by finding the cubic root of the sum of each time-frequency representation.
- The DCT is then applied to these values to provide 32 cepstral features.

In this study, 32 GTCC values are used for each ear measured at center frequencies from 0 to 20 kHz. Additionally, the log energy of each signal is also included with each GTCC measurement resulting in a total of 33 values. Because these are measured for both ears, the final vector length for GTCC cues is 66.

3 MULTI-FEATURE ML PERFORMANCE FOR SSL TASKS

Most ML algorithms for SSL tasks are trained using a combination of features. In this section, an examination of multi-feature training algorithms is presented in two separate HRTF conditions. The first is referred to as the matched condition in which the HRTF set used for testing is the same as that used in the training stage. The second is referred to as the mismatched condition in which the algorithm is trained on data created using one HRTF set and tested on data created by another. The authors propose a new feature combination of CCF, ILD, and GTCCs and evaluate its performance alongside two other current methods referred henceforth as the Ma [22] and Wu [25] methods. These methods are described in detail in the successive sections.

3.1 ML SSL Methods

In this study, results are compared with two state-of-the-art SSL algorithms. The first is that presented by Ma et al. in [22], which uses CCF and ILD cues to perform SSL in the horizontal plane with a high degree of accuracy. The second is a 3D SSL algorithm presented by Wu and Talagala in [25]. This study shows how Interaural Phase Difference (IPD) and ILD cues are effective in training ML algorithms to determine both the horizontal and vertical positions of a sound source. These algorithms and their implementation are described in greater detail below.

3.1.1 DNN Training Using CCF and ILD, the Ma Method

Interaural cues, ITD, and ILD are the most commonly used cues in binaural source localization models. However, Ma et al. [22] found that the normalized CCF contains more information relating to source localization than the ITD. Ma's work showed that the performance of the CCF can be further improved by combining it with a single value ILD measurement taken as an average ILD value across the entire frequency range. In Ma's work, the binaural signals are filtered by a Gammatone filter-bank to obtain several sub-bands, each with their own measurements of CCF and ILD. For each sub-band, a DNN was trained with the source azimuth determined by the average estimation from each DNN. An accuracy upward of 90% was found using this method when estimating the horizontal location of a sound source to one of 72 evenly spaced source positions in four separate reverberant environments. This method was shown to be accurate in the presence of multiple sound sources.

Because the authors are interested in only locating a single sound source in this study, a single CCF and ILD measurement is taken across the entire frequency range and used to train a single DNN. Although Ma's method is a robust method for localization in the horizontal, it is not intended specifically to be used for vertical localization. It should also be noted that the experiments performed in that study were performed using a sampling rate of 16 kHz, meaning that the CCF of length ± 1 ms had a length of 33. This also means that the frequency range of the stimuli was restricted to 8 kHz.

When replicating this method with a higher sampling rate, several adjustments needed to be made. The Ma method uses a CCF with a lag range of ± 1 ms, producing an 89-dimensional binaural feature combined with a single ILD value measurement. The resultant vector is of length 90 and is used to train a DNN with two hidden layers. The number of hidden layer nodes in the replication of the Ma method is increased from 128 to 300 because of the increase in the length of CCF when using a larger sampling rate. The network performs classification with a "softmax" activation function applied to the output layer.

3.1.2 Random Forest Training Using ILD and IPD, the Wu Method

In 2019, Wu and Talagala [25] presented a binaural SSL model capable of localizing both the azimuth and elevation of a single sound source. This work combined the interaural features ILD and IPD and used Random Forest (RF) classification. IPD measures the difference in the phase of a wave that reaches each ear. The advantage of using a random forest over a neural network is that they can use information gain to determine which features are best for splitting the data into subsequent classifications. Wu's method takes advantage of this by training the algorithm on the phase and magnitude differences of a full Fast Fourier Transform (FFT) of the binaural signal. ILD and IPD values were extracted using Hamming windows of 16-ms length with an 8-ms overlap. Their study used a sampling rate of 16 kHz with an FFT of size 512, resulting in 256 values for IPD and ILD ranging from 0 to 8 kHz.

To replicate this method, IPD and ILD can be calculated from a binaural signal, respectively as v_f^p and v_f^m using the following formulae:

$$v_f^p = \angle\left|\frac{X_{l,k,f}}{X_{r,k,f}}\right|, \tag{5}$$

$$v_f^m = 20\log_{10}\left|\frac{X_{l,k,f}}{X_{r,k,f}}\right|, \tag{6}$$

where k and f represent the frame and frequency index and X_l and X_r are the FFTs of the left and right binaural signals, respectively. Eq. (6) shows a measurement of ILD similar to Eq. (3). However, in Eq. (3), an average ILD is taken from within the frequency range between 1.5 and 20 kHz. For the Wu method, all 256 ILD measurements are used to train

the network. Although this method was shown to outperform the probabilistic piecewise affine mapping proposed by Deleforge in [26], it was not tested for performance in the mismatched HRTF condition.

The Wu method uses a combination of ILD and IPD values extracted from the magnitude and phase components of the FFT of a binaural signal. A 512-point FFT size is used in this method resulting in 256 values for both ILD and IPD values measured between 0 and 8 kHz. This feature space is fed into two separate random forest algorithms, one for determining the source azimuth and the other for determining the elevation. The RF algorithm for azimuth estimation consists of 50 trees and a maximum depth of 32, whereas the one for elevation estimation uses 100 trees and a maximum depth of 64. Both algorithms use information gain to determine feature importance and reduce the number of individual ILD and IPD points used to train the network. In replicating this method, the maximum depth of each tree in Wu's method is increased from 32 to 64 for horizontal plane localization and from 64 to 128 for vertical plane localization.

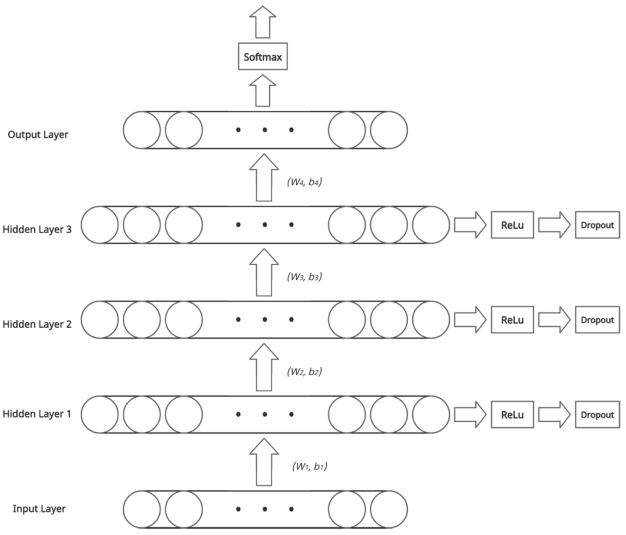


Fig. 1. The architecture of the Deep Neural Network (DNN) used to perform sound source localization. ReLu = rectified linear unit.

3.2 Comparing Multi-Feature Localization Performance

The methods employed by both Ma and Wu can be seen as benchmarks for localization tasks using ML. Implementing their methods provides a reference and also allows for examination of the two methods in the mismatched HRTF scenario—something neither study investigated. A new multi-feature array for SSL comprised of a combination of CCF, ILD, and GTCC cues is proposed. This array can be seen as an extension of the Ma method with an additional spectral cue in GTCCs. Although both the Ma and Wu methods were originally established using a sampling rate of 16 kHz, in this work, a consistent sampling rate of 44.1 kHz is used to compare methods. Thus, these methods are being applied to a larger frequency range, and the size of the input vectors for each network is increased.

The combined feature set of CCF, ILD, and GTCCs of "the proposed method" is of length 158 and is used to train both a DNN and an RF algorithm. When training the DNN with this feature set, the same architecture is used as shown in Fig. 1 using a hidden layer size of 300 nodes. DNN training is performed in python using Keras [27] as a high-level API to run a Tensorflow backend [28]. There are a total of 62,500 synthesized binaural signals per HRIR, and these were split into groups of 80% for training and 20% for testing, meaning that the algorithm was trained on a dataset of size 50,000 and tested on a dataset of size 12,500.

The input layer is equal to the length of the input feature, whereas the output layer is equal to 25, the number of classes in this classification task. Each one of these classes corresponds to a location along either the horizontal plane, median plane, or sagittal plane at 45°. There are three hidden layers consisting of a variable number of nodes, which use a rectified linear unit (ReLU) activation function. ReLU activation is a piecewise linear function that will output the input directly if it is positive or will output 0 if it is negative. Following each hidden layer is a Dropout layer in which 20% of the nodes are randomly removed to avoid overfitting. The DNN training algorithm uses the "Adam" optimizer [29], with a learning rate of $10e^{-4}$. The output layer uses "softmax" activation to select an output class. The network is compiled using "categorical crossentropy" as its loss function. The network is trained for a total of 300 epochs with early stopping employed if there is no change in model performance for 10 epochs. Mini-batch gradient descent is used with batch sizes of 200.

When training the RF algorithm with this feature set, 50 trees with a max depth of 32 is used for horizontal estimation, and100 trees with a maximum depth of 64 is used for vertical localization. These two new implementations of these algorithms are compared with the Ma and Wu methods in terms of performance in the horizontal plane, median plane, and sagittal plane at 45°. The average result of these algorithms along these planes can be seen in Figs. 2–4.

These results show that localization in the matched HRTF condition consistently outperforms localization in the mismatched HRTF condition. This result is in line with most ML algorithms, which perform better when tested on the same data that they were trained on. With this in mind, the authors are more interested in the results from the mismatched HRTF condition because they are not prone to overfitting. Here, the performance among the algorithms varies depending on the localization plane. Although the proposed feature set implemented using a DNN performs worst along the horizontal plane, it outperforms the other algorithms in both elevation estimation tasks. The proposed feature set implemented using the RF algorithm performs on a comparable level of success to both the Wu and Ma methods in terms of horizontal localization. In Fig. 5, the accuracy of each HRTF for predicting source elevation along the sagittal plane at 45° can be visualized when testing against each of the 45 HRTF sets in the CIPIC database.

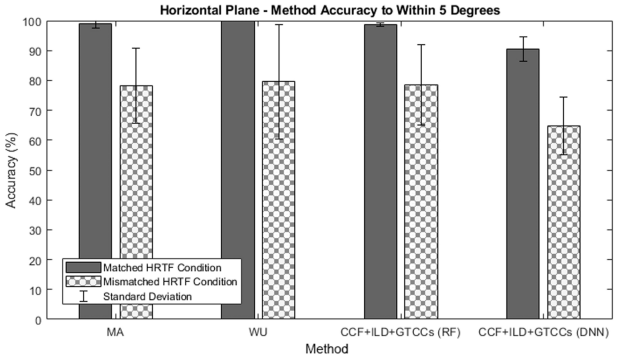


Fig. 2. Horizontal plane accuracy in the matched and mismatched condition for azimuth estimation to within 5°. CCF = cross-correlation function; DNN = Deep Neural Network; GTCC = Gammatone Filter Cepstral Coefficient; HRTF = Head-Related Transfer Function; ILD = interaural sound level difference; RF = Random Forest.

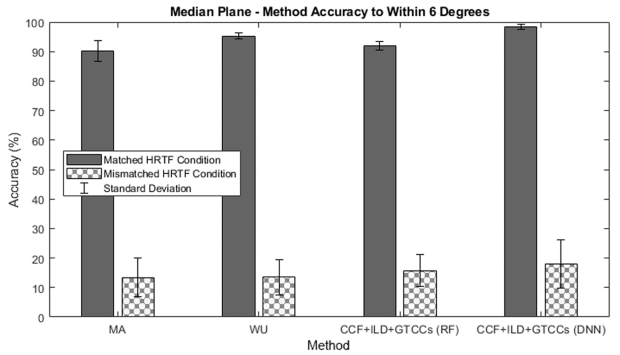


Fig. 3. Median plane accuracy in the matched and mismatched condition for elevation estimation to within 6°. CCF = cross-correlation function; DNN = Deep Neural Network; GTCC = Gammatone Filter Cepstral Coefficient; HRTF = Head-Related Transfer Function; ILD = interaural sound level difference; RF = Random Forest.

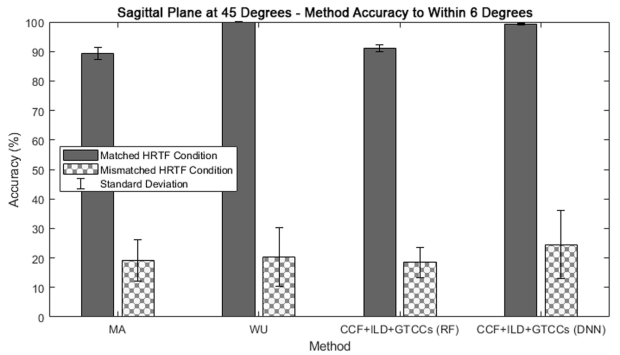


Fig. 4. Sagittal plane at 45° accuracy in the matched and mismatched condition for elevation estimation to within 6°. CCF = cross-correlation function; DNN = Deep Neural Network; GTCC = Gammatone Filter Cepstral Coefficient; HRTF = Head-Related Transfer Function; ILD = interaural sound level difference; RF = Random Forest.

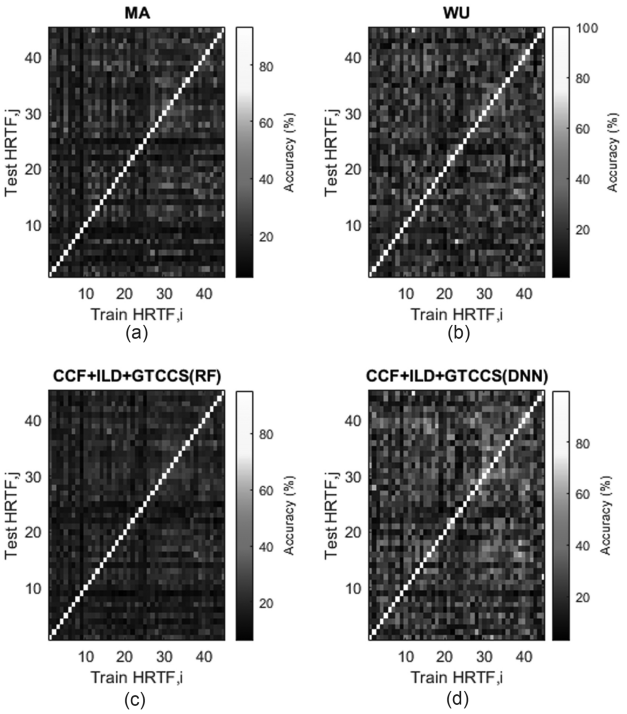


Fig. 5. Sagittal plane at 45° method performance for estimating elevation to within 6°. A single training HRTF is used to train the algorithm and subsequently tested on each of the 45 HRTFs from the CIPIC dataset. Two popular methods of binaural localization are compared to two new proposed methods. The proposed method results are shown in (d). CCF = cross-correlation function; DNN = Deep Neural Network; GTCC = Gammatone Filter Cepstral Coefficient; HRTF = Head-Related Transfer Function; ILD = interaural sound level difference; RF = Random Forest.

4 HRTF CLUSTERING

In Fig. 5, the performance of each method can be seen as a matrix depicting localization accuracy for each algorithm trained on a single HRTF, i, and tested on a subsequent HRTF, j. The results of the proposed method, which combines CCF, ILD, and GTCC cues and a DNN algorithm, are shown in the bottom-right of Fig. 5. It should be noted that although the highest performance accuracy can be seen in the matched HRTF condition, there are many instances of accuracy beyond 50% observed in the mismatched HRTF condition. In many cases, these performances can also be observed to be commutable; for instance, this is the case for HRTFs 12 and 45, which were both measured using a KEMAR head microphone but with different ears. The proposed method found an accuracy of 86.3% when training on HRTF 45 and testing on HRTF 12 and 83.6% when training on HRTF 12 and testing on HRTF 45. With this in mind, a clustering method is proposed to group HRTFs together into groups in which a higher-than-usual localization accuracy is observed in the mismatched HRTF condition.

In this study, a clustering analysis based on the Affinity Propagation (AP) algorithm [30] is employed to cluster similar HRTF sets together using the results shown in Fig. 5(d) as a Similarity Matrix (SM). The AP al-

gorithm has been shown to provide a robust engineering solution to clustering problems in audio signal processing, such as speaker clustering [31] and acoustic control of crosstalk [32] and in VANETs for mobile telecommunications [33].

Compared with other common clustering methods, such as K-means and hierarchical clustering, the AP algorithm has many unique advantages. For instance, unlike K-means, the AP algorithm does not need any prior information on the number of clusters and will determine the most appropriate number of clusters from the data. The AP algorithm also selects an actual measured HRTF to define a cluster center, whereas K-means clustering selects a center relating to an average of all sample points, which most likely will not correspond to a real HRTF. Furthermore, the AP algorithm can be applied to problems in which the SM is not symmetrical, i.e., the performance of the DNN trained on HRTF set i and tested on HRTF set j is not equal to the performance of the DNN trained on HRTF set j and tested on HRTF set i.

The AP algorithm was used by Wang et al. [34] to perform HRTF clustering based on horizontal localization performance only. Building upon this work, here, an Adaptive form of the AP algorithm, AAP, [30] is used to cluster HRTF sets together based on performance in the vertical plane.

The AP algorithm is iterative and uses a damping factor, λ, between the values of 0 and 1, which must be selected prior to commencing the algorithm. An ideal value for λ enhances convergence and dampens oscillations. In the AAP algorithm, the value of λ is adapted using a moving window observing the iterations and under the competing influences of damping oscillatory tendencies by $\lambda \rightarrow 1$ and faster convergence with $\lambda \rightarrow 0$. This stabilizing adaptation of λ accelerates convergence.

The *Silhouette Index* (SI) is a measure of the compactness and separation of clusters. It is an internal validation index because it provides a measure of goodness of the clustering without recourse to external information. In the AAP algorithm, it is used to estimate the optimum number of clusters. SI is the average of all Silhouette values that lie between –1 and 1. High Silhouette values indicate that an HRTF is well matched to the cluster it is in and poorly matched to neighboring clusters. A low or negative SI value would indicate that the clustering configuration has too many or too few clusters. A more in-depth explanation of Silhouette values can be found in [35].

4.1 Clustering Analysis and Performance

Because the primary concern of this study is to improve the accuracy of vertical localization, HRTF set clustering was performed with respect to the proposed feature set of CCF, ILD, and GTCCs cues and its performance along the sagittal plane. Because this feature set contains CCF and ILD cues that generalize well to localization tasks along the horizontal [36, 37, 22]; it is believed that after clustering, performance along the horizontal plane will still be sufficiently accurate. Performing clustering with the AAP algorithm reveals an optimum clustering of 4, given the SM found in Fig. 5.

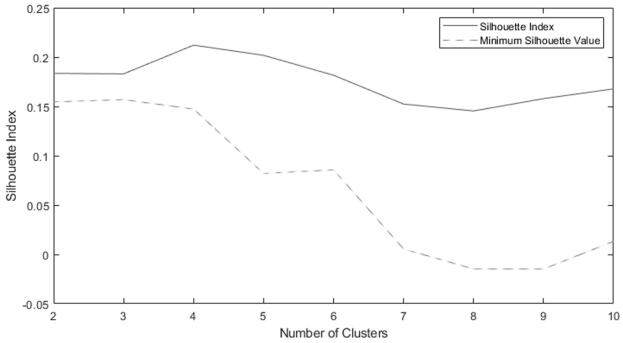


Fig. 6. The Silhouette Index values for two to ten clusters found using the Adaptive Affinity Propagation algorithm.

These clusters range in size from 9 to 13 and can be seen in Table 1. The SI for this particular clustering is 0.2121, found using a damping factor, λ, of 0.5. This value was the highest SI value for arrangements from two clusters up to ten. Although the optimum number of clusters is comparatively smaller than results produced in other clustering studies such as [34], which determined an optimum clustering of 7, it should be noted how abstract this clustering is because it is based on localization scores from binaural audio signals. In [34], clustering is based on azimuthal localization scores, whereas here, clustering is based on localization along the sagittal plane at 45°. SI values and minimum Silhouette values for each cluster number can be seen in Fig. 6.

Following this clustering analysis, sagittal plane localization is repeated at 45° with respect to elevation angle. This time HRTF clustering is employed, whereby the four central HRTF sets from each cluster are used to train the DNN. This network is then tested on the 45 HRTF conditions, whereby four are matched and 41 are mismatched. The result of this test shows that clustering reduces performance in the matched condition while enhancing it in the mismatched condition as shown in Fig. 7. Here, performance goes from 99.27% in the matched condition to 92.92% in the matched condition with no central HRTF performing greater than 94.8%. In the mismatched condition, performance increases from 24.54% to 45.8%. This improvement of over 20% in the mismatched condition shows a much better generalization in the network to mismatched HRTF conditions. Training the DNN using a single HRTF set condition takes on average 13.96 s, whereas training the DNN using four HRTF sets as is the case in this clustered approach takes on average 45.93 s.

4.2 Spherical Localization and Angular Error

When calculating the error of source estimation on the unit sphere, an angular error that reflects the error in 3D space must be used. In order to perform this calculation of angular error, both ground truth source angles and predicted source angles must first be converted from degrees to radi-

Table 1. Clustering result of Head-Related Transfer Function (HRTF) sets from the CIPIC database with four clusters.

Cluster index	No. of HRTF sets	HRTF indices	Central HRTF
Cluster 1	13	1,3,7,10,11,13,15,17,19,20,32,38,42	1
Cluster 2	13	4,5,8,12,14,16,18,27,29,39,40,44,45	27
Cluster 3	10	2,6,9,23,24,26,28,36,41,43	28
Cluster 4	9	21,22,25,30,31,33,34,35,37	37

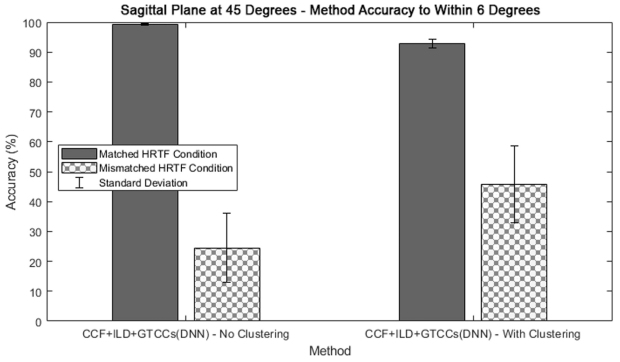


Fig. 7. The performance of the proposed feature set along the Sagittal Plane at 45°, correctly classified to within 6°. Performance is compared for training on a single Head-Related Transfer Function (HRTF) set versus training on four central HRTF sets. CCF = cross-correlation function; DNN = Deep Neural Network; GTCC = Gammatone Filter Cepstral Coefficient; ILD = interaural sound level difference; RF = Random Forest.

ans. Following this, the 3D coordinates must be converted from polar to cartesian coordinates subbing in a constant value for distance, which, in this study, is defined as 1 m. This angular error can be defined by the absolute angular difference between two directional vectors in a Cartesian coordinate system and calculated as

$$\epsilon = \arccos < \frac{d_{(\theta,\phi,r)} . \hat{d}_{(\theta,\phi,r)}}{|d_{(\theta,\phi,r)}||\hat{d}_{(\theta,\phi,r)}|} >, \tag{7}$$

where $d_{(\theta,\ \phi,\ r)}$ and $\hat{d}_{(\theta,\phi,r)}$ are the ground truth and estimated source directions, respectively. The angular error can then be converted back from radians to degrees.

4.3 Spherical Localization Performance

Given the feature set proposed for this method of performing SSL, spherical localization is performed using a combination of two ML algorithms. The first is a random forest network trained to estimate the source azimuth. It consists of 50 trees and has a maximum depth of 32 features. The second algorithm is a DNN used to estimate the source elevation. This DNN has three hidden layers consisting of 300 nodes, a dropout rate of 20% following each hidden layer and a learning rate of $10e^{-4}$. It also uses a ReLU activation function.

The performance of the proposed method is presented with and without HRTF clustering. Binaural signals are generated by convolving HRTF datasets from the CIPIC database [15] with random speech signals from the TSP speech dataset [16]. Here, the binaural signals are received from 625 source locations in the front hemifield with 25 sagittal planes each containing 25 HRTF measurements along the vertical. A total of 100 randomly selected speech samples from the TSP database were used to create 100 synthesized binaural signals per HRIR resulting in 62,500 synthesized signals, 80% of which were used for training and 20% used for testing. Performance is measured in terms of average angular error in both the matched and mismatched HRTF condition. White gaussian noise is added to the speech signals prior to feature extraction at four different signal-to-noise ratios (SNRs): 10 dB, 20 dB, 30 dB, and No Noise. For each SNR, a new dataset of the same size was created. These results can be seen in Table 2.

These results clearly show that HRTF clustering significantly decreases angular error in the mismatched HRTF condition across all SNR levels. For all SNR levels the results found using HRTF clustering are greater than 5° better than those found without HRTF clustering. Although performance using HRTF clustering is worse in the matched condition, it is important to note that there are four matched

Table 2. Spherical localization—average angular error in noisy environments with and without HRTF clustering.

	Average Angular Error							
SNR	10 dB		20 dB		30 dB		No Noise	
HRTF condition	Matched	Mismatch	Matched	Mismatch	Matched	Mismatch	Matched	Mismatch
No clustering	7.04°	19.7°	5.1°	18.72°	3.92°	17.65°	2.55°	16.87°
With clustering	9.37°	13.8°	7.35°	11.79°	6.17°	10.6°	5.22°	10.23°

HRTF = Head-Related Transfer Function; SNR = signal-to-noise ratio.

Table 3. Spherical localization—average azimuth accuracy with and without HRTF clustering.

Azimuth Accuracy Comparison, Error tolerance to ≤5°								
SNR	10 dB		20 dB		30 dB		No Noise	
HRTF Condition	Matched	Mismatch	Matched	Mismatch	Matched	Mismatch	Matched	Mismatch
No Clustering	96.252%	68.5%	97.74%	75.84%	97.156%	79.72%	98.752%	80.34%
With Clustering	93.01%	85.32%	93.38%	85.42%	93.59%	85.36%	93.48%	85.57%

HRTF = Head-Related Transfer Function; SNR = signal-to-noise ratio.

conditions and 41 mismatched conditions. And although there is a trade-off between matched and mismatched performance using HRTF clustering, it can be concluded that this is worth taking given the improvement of results in a majority of HRTF conditions.

In Tables 3 and 4, the azimuth and elevation accuracy of the algorithms are shown. These are categorized as the percentage of samples correctly classified to within 5° in the case of azimuthal accuracy and within 6° in the case of elevation accuracy. These limits were selected because the smallest increments in horizontal HRTF measurements in the CIPIC database are 5° and in 6° for vertical HRTF measurements. In terms of Azimuthal accuracy, these results show a consistency in results when using HRTF clustering with performance remaining approximately equal across all noise conditions in both the matched and mismatched conditions. This is not the case when HRTF clustering is not applied, however, because although performance is greater than with HRTF clustering in the matched condition, it is increasingly worse in the mismatched HRTF condition with increasing levels of noise.

When examining elevation accuracy, it can be seen that results both with and without HRTF clustering diminish with increasing levels of noise in the signal. Performance in the matched condition is consistently better across all noise levels when there is no HRTF clustering. However, mismatched HRTF performance is better in all noise conditions when HRTF clustering is applied with an accuracy of 28.77% at SNR = 10 dB better than an accuracy of 25.87% in the no noise condition when there is no HRTF clustering. Mismatched performance is improved by up to 13% when HRTF clustering is employed.

5 CONCLUSION

In this study, the AAP clustering algorithm is used to cluster similar HRTF sets from the CIPIC database together. This clustering is based on the performance of the proposed feature set of CCF, ILD, and GTCC cues in estimating sound source elevation on the sagittal plane at 45° using a DNN. This clustering algorithm reveals four unique clusters ranging in size from nine to 13 HRTF sets, each with their own central HRTF that is most representative of the entire cluster. Using these four HRTF sets, a DNN is trained and used to perform the same localization task along the front hemifield from −80° to +80°.

In Fig. 7, the results of a localization experiment on the sagittal plane at 45° are shown comparing matched and mismatched HRTF conditions with or without the use of HRTF clustering. This experiment featured no additional noise on the binaural signal. Here, it is shown that the use of HRTF clustering improves the performance of vertical source estimation by over 20% in the mismatched HRTF condition. Concurrently, HRTF clustering also decreases localization performance in the matched HRTF by approximately 6%. However, the matched HRTF condition accounts for less than 9% of all test conditions.

When examining the effect of HRTF clustering on spherical localization, it was shown that even in noisy environments, HRTF clustering improves the performance of the SSL algorithm in the mismatched HRTF condition. At an SNR of 10 dB, the average angular error of the algorithm using HRTF clustering is 13.8° which is superior to the value of 16.87° exhibited in the condition where no clustering is used and no noise is present. On examination of error with regards to azimuth and elevation, HRTF clustering

Table 4. Spherical localization—average elevation accuracy with and without HRTF clustering.

Elevation Accuracy Comparison, Error tolerance to ≤6°								
SNR	10 dB		20 dB		30 dB		No Noise	
HRTF Condition	Matched	Mismatch	Matched	Mismatch	Matched	Mismatch	Matched	Mismatch
No Clustering	79.13%	20.11%	84.38%	23.1%	89.775%	25.16%	94.64%	25.87%
With Clustering	64.13%	28.77%	73.08%	33.73%	80.19%	35.67%	87.5%	38.75%

HRTF = Head-Related Transfer Function; SNR = signal-to-noise ratio.

results in a much more robust performance in the mismatched HRTF condition for both. Azimuthal performance is shown to be robust across all levels of noise while elevation performance surpasses no clustering performance by up to 13% in the mismatched HRTF condition. From these results it is evident that HRTF clustering can vastly improve the robustness of binaural SSL to unseen HRTF conditions.

6 ACKNOWLEDGMENT

This research was supported by Science Foundation Ireland Investigator Award 13/IA/1900 and the Trinity College Dublin Sigmedia Research Group. The authors acknowledge the helpful discussions with their colleague Dr. Enda Bates and the support of the ADAPT SFI Centre for Digital Media Technology funded by Science Foundation Ireland.

7 REFERENCES

[1] L. Rayleigh, "XII. On Our Perception of Sound Direction," *Lond. Edinb. Dublin Philos. Mag. J. Sci.*, vol. 13, no. 74, pp. 214–232 (1907 Feb). http://dx.doi.org/10.1080/14786440709463595.

[2] W. M. Hartmann, B. Rakerd, Z. D. Crawford, and P. X. Zhang, "Transaural Experiments and a Revised Duplex Theory for the Localization of Low-Frequency Tones," *J. Acoust. Soc. Am.*, vol. 139, no. 2, pp. 968–985 (2016 Feb.). https://doi.org/10.1121/1.4941915.

[3] B. A. Edmonds and K. Krumbholz, "Are Interaural Time and Level Differences Represented by Independent or Integrated Codes in the Human Auditory Cortex?" *J. Assoc. Res. Otolaryngol.*, vol. 15, no. 1, pp. 103–114 (2013 Nov.). https://doi.org/10.1007/s10162-013-0421-0.

[4] J. C. Middlebrooks, "Sound Localization," in M. J. Aminoff, F. Boller, and D. F. Swaab (Eds.), *Handbook of Clinical Neurology*, vol. 129, pp. 99–116 (Elsevier, Amsterdam, Netherlands, 2015). http://dx.doi.org/10.1016/B978-0-444-62630-1.00006-8.

[5] C. C. Pratt, "The Spatial Character of High and Low Tones." *J. Exp. Psychol.*, vol. 13, no. 3, pp. 278–285 (1930 Jun). http://dx.doi.org/10.1037/h0072651.

[6] V. C. Raykar, R. Duraiswami, and B. Yegnanarayana, "Extracting the Frequencies of the Pinna Spectral Notches in Measured Head Related Impulse Responses," *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 364–374 (2005 Jul.). http://dx.doi.org/10.1121/1.1923368.

[7] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, "Localization Using Nonindividualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 94, no. 1, pp. 111–123 (1993 Jul.). http://dx.doi.org/10.1121/1.407089.

[8] K. Inanaga, Y. Yamada, and H. Koizumi, "Headphone System With Out-of-Head Localization Applying Dynamic HRTF (Head-Related Transfer Function)," presented at the *98th Convention of the Audio Engineering Society* (1995 Feb.), paper 4011.

[9] C. I. Cheng and G. H. Wakefield, "Introduction to Head-Related Transfer Functions (HRTFs): Representations of HRTFs in Time, Frequency, and Space," presented at the *107th Convention of the Audio Engineering Society* (1999 Sep.), paper 5026.

[10] A. Härmä, J. Jakka, M. Tikander, et al., "Augmented Reality Audio for Mobile and Wearable Appliances," *J. Audio Eng. Soc.*, vol. 52, no. 6, pp. 618–639 (2004 Jun.).

[11] P. Jain, J. Manweiler, and R. R. Choudhury, "Overlay: Practical Mobile Augmented Reality," in *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 331–344 (Florence, Italy) (2015 May). http://dx.doi.org/10.1145/2742647.2742666.

[12] F. Keyrouz, "Advanced Binaural Sound Localization in 3-D for Humanoid Robots," *IEEE Trans. Instrum. Meas.*, vol. 63, no. 9, pp. 2098–2107 (2014 Sep.). http://dx.doi.org/10.1109/TIM.2014.2308051.

[13] H. Nakashima and T. Mukai, "3D Sound Source Localization System Based on Learning of Binaural Hearing," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, pp. 3534–3539 (Waikoloa, HI) (2005 Oct.). http://dx.doi.org/10.1109/ICSMC.2005.1571695.

[14] Y. Zhang, G. Liu, and B. Luo, "Finite-Time Cascaded Tracking Control Approach for Mobile Robots," *Inform. Sci.*, vol. 284, pp. 31–43 (2014 Nov.). http://dx.doi.org/10.1016/j.ins.2014.06.037.

[15] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF Database," in *Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102 (New Paltz, NY) (2001 Oct.). http://dx.doi.org/10.1109/ASPAA.2001.969552.

[16] P. Kabal, *TSP Speech Database* (McGill, Quebec, Canada, 2002).

[17] J. S. Garofalo, L. F. Lamel, W. M. Fisher, et al., "The DARPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus CD-ROM," *NISTIR 4930* (1993 Feb.).

[18] K. Iida, M. Itoh, A. Itagaki, and M. Morimoto, "Median Plane Localization Using a Parametric Model of the Head-Related Transfer Function Based on Spectral Cues," *Appl. Acoust.*, vol. 68, no. 8, pp. 835–850 (2007 Aug.). http://dx.doi.org/10.1016/j.apacoust.2006.07.016.

[19] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1984).

[20] R. F. Lyon, *Human and Machine Hearing: Extracting Meaning From Sound* (Cambridge University Press, Cambridge, UK, 2017). http://doi.org/10.1017/9781139051699.

[21] J. Blauert, *Räumliches Hören* (Hirzel, Stuttgart, Germany, 1974).

[22] 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 Trans. Audio Speech Lang. Process.*, vol. 25, no. 12, pp. 2444–2453 (2017 Dec.). http://dx.doi.org/10.1109/TASLP.2017.2750760.

[23] X. Zhao, Y. Shao, and D. Wang, "CASA-Based Robust Speaker Identification," *IEEE Trans. Audio Speech*

Lang. Process., vol. 20, no. 5, pp. 1608–1616 (2012 Jul.). http://dx.doi.org/10.1109/TASL.2012.2186803.

[24] X. Valero and F. Alias, "Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification," *IEEE Trans. Multimed.*, vol. 14, no. 6, pp. 1684–1689 (2012 Dec.). http://dx.doi.org/10.1109/TMM.2012.2199972.

[25] X. Wu, D. S. Talagala, W. Zhang, and T. D. Abhayapala, "Individualized Interaural Feature Learning and Personalized Binaural Localization Model," *Appl. Sci.*, vol. 9, no. 13, paper 2682 (2019 Jun.). http://dx.doi.org/10.3390/app9132682.

[26] A. Deleforge, F. Forbes, and R. Horaud, "Acoustic Space Learning for Sound-Source Separation and Localization on Binaural Manifolds," *Int. J. Neural Syst.*, vol. 25, no. 1, paper 1440003 (2014 Apr.). http://dx.doi.org/10.1142/S0129065714400036.

[27] F. Chollet, "Keras," https://keras.io/. (accessed Mar. 27, 2015).

[28] M. Abadi, P. Barham, J. Chen, et al., "Tensorflow: A System for Large-Scale Machine Learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, pp. 265–283 (Savannah, GA) (2016 Nov.).

[29] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv preprint arXiv:1412.6980* (2014 Dec.).

[30] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo, "Adaptive Affinity Propagation Clustering," *arXiv preprint arXiv:0805.1096* (2008 May).

[31] X. Zhang, J. Gao, P. Lu, and Y. Yan, "A Novel Speaker Clustering Algorithm via Supervised Affinity Propagation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4369–4372 (Las Vegas, NV) (2008 Apr.). http://dx.doi.org/10.1109/ICASSP.2008.4518623.

[32] K.-S. Lee, "Position-Dependent Crosstalk Cancellation Using Space Partitioning," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 6, pp. 1228–1239 (2013 Jun.). http://doi.org/10.1109/TASL.2013.2248713.

[33] H. Shahwani, T. D. Bui, J. P. Jeong, and J. Shin, "A Stable Clustering Algorithm Based on Affinity Propagation for VANETs," in *Proceedings of the 19th International Conference on Advanced Communication Technology (ICACT)*, pp. 501–504 (PyeongChang, South Korea) (2017 Feb.). http://doi.org/10.23919/ICACT.2017.7890140.

[34] J. Wang, J. Wang, K. Qian, X. Xie, and J. Kuang, "Binaural Sound Localization Based on Deep Neural Network and Affinity Propagation Clustering in Mismatched HRTF Condition," *EURASIP J. Audio Speech Music Process.*, vol. 2020, no. 1, paper 4 (2020 Feb.). http://dx.doi.org/10.1186/s13636-020-0171-y.

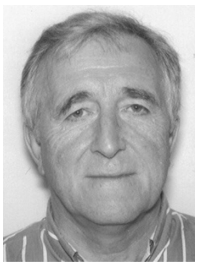
[35] A. Starczewski and A. Krzyżak, "Performance Evaluation of the Silhouette Index," in L. Rutkowski, M. Korytkowski, and R. Scherer (Eds.), *Artificial Intelligence and Soft Computing*, Lecture Notes in Computer Science, vol. 9120, pp. 49–58 (Springer, Cham, Switzerland, 2015). http://dx.doi.org/10.1007/978-3-319-19369-4_5.

[36] X. Wan and Z. Wu, "Sound Source Localization Based on Discrimination of Cross-Correlation Functions," *Appl. Acoust.*, vol. 74, no. 1, pp. 28–37 (2013 Jan.). http://dx.doi.org/10.1016/j.apacoust.2012.06.006.

[37] J. C. Murray, H. Erwin, and S. Wermter, "Robotics Sound-Source Localization and Tracking Using Interaural Time Difference and Cross-Correlation," in *Proceedings of the AI Workshop on NeuroBotics*, pp. 89–97 (Ulm, Germany) (2004 Sep.).

THE AUTHORS

Hugh O'Dwyer



Frank Boland

Hugh O'Dwyer is a researcher at the Sigmedia lab at Trinity College Dublin, where he recently completed his Ph.D. thesis entitled *Sound Source Localization and Virtual Testing of Binaural Audio*. His research interests include binaural audio, spatial audio, and psychoacoustics. Prior to his postgraduate studies, he completed his undergraduate degree in Biomedical Engineering at the same university. He has published works on topics including ambisonic microphones, virtual headphone testing, and sound source localization using machine learning algorithms. He also works as a sound engineer and music producer out of several recording studios in Dublin.

•

Francis Boland is Emeritus Professor of Engineering Science at the Department of Electronic and Electrical Engineering at Trinity College Dublin. He graduated as an electrical engineer from University College Dublin, Bachelor of Engineering, and then joined Sheffield Polytechnic as a research assistant on a joint project with the British Steel Corporation Special Steels Research Laboratories. He completed his Ph.D. thesis on mathematical models of the metallurgical process with a subsequent publication awarding him the Kelvin Premium by the Institution of Electrical Engineers. His recent research has included immersive audio over headphones and the development of the Thrive IP, which was licensed by Google in 2015 and provides a basis for the open-source environment Resonance Audio.