Hearing Research xxx (xxxx) 108884

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research Paper

Localizing concurrent sound sources with binaural microphones: A simulation study

Jakeh Orr *, William Ebel, Yan Gai

School of Science and Engineering, Saint Louis University, Saint Louis, MO 63105, USA

ARTICLE INFO

Article history:
Received 30 May 2023
Received in revised form 4 September 2023
Accepted 9 September 2023

Keywords:
Sound localization
HRTF
ITD
ILD
Robotics
Sparseness
Reverberations

ABSTRACT

The human auditory system can localize multiple sound sources using time, intensity, and frequency cues in the sound received by the two ears. Being able to spatially segregate the sources helps perception in a challenging condition when multiple sounds coexist. This study used model simulations to explore an algorithm for localizing multiple sources in azimuth with binaural (i.e., two) microphones. The algorithm relies on the "sparseness" property of daily signals in the time-frequency domain, and sound coming from different locations carrying unique spatial features will form clusters. Based on an interaural normalization procedure, the model generated spiral patterns for sound sources in the frontal hemifield. The model itself was created using broadband noise for better accuracy, because speech typically has sporadic energy at high frequencies. The model at an arbitrary frequency can be used to predict locations of speech and music that occurred alone or concurrently, and a classification algorithm was applied to measure the localization error. Under anechoic conditions, averaged errors in azimuth increased from 4.5° to 19° with RMS errors ranging from 6.4° to 26.7° as model frequency increased from 300 to 3000 Hz. The low-frequency model performance using short speech sound was notably better than the generalized cross-correlation model. Two types of room reverberations were then introduced to simulate difficult listening conditions. Model performance under reverberation was more resilient at low frequencies than at high frequencies. Overall, our study presented a spiral model for rapidly predicting horizontal locations of concurrent sound that is suitable for real-world scenarios.

1. Introduction

Human listeners can localize more than one sound source concurrently (Zhong and Yost, 2017, Keller and Takahashi, 2005). Spatial separations between speech and interfering noise can often improve speech perception, a phenomenon called "spatial release from masking" (Saberi et al., 1991). However, the ability for wearers of conventional hearing aids to localize sound sources is poor because of compromised localization cues (Loiselle et al., 2016). Sound-processing algorithms that can automatically decode sound locations and subsequently segregate the sound into different streams according to the locations would be highly desirable for hearing devices. The goal of this paper is to explore a fast and accurate localization algorithm that is invariant with common sound types in our daily life.

Although an array of directional microphones can enhance speech perception (Saunders, 1997), users typically prefer wearing more discreet devices with binaural (i.e., two) microphones. With two microphones, the most successful underdetermined blind-source separation relies on the *sparseness* of sound signals. Fig. 1 shows the example used in one of our simulations, having three concurrent sound sources located at different places. Here, a guitar is placed 80° left to the front center, a female speaker is placed 30° left to the front center, and a male speaker is placed 80° right to the front center. Although sound sources in our daily life often overlap in the time domain (i.e., "concurrent"), when projected into a two-dimensional time-frequency domain, the spectrograms rarely occupy the same time and frequency spots, as shown in the illustration of Fig. 2B. In other words, adding the frequency dimension is crucial to the sparseness property. With each sound source carrying distinct spatial cues, different time-frequency points can be classified into separate source locations. Signals can then be extracted with a time-frequency binary mask if sound segregation is also desired (Makino et al., 2007).

According to the duplex theory (Mills, 1972, Tollin, 2009), at low sound frequencies (e.g., < 2 kHz), sound localization in azimuth is based on interaural time differences (ITDs) caused by differences in the sound's arrival time at the two ears. At high frequencies (e.g., > 2 kHz), interaural level differences (ILDs) caused by head shadows dominate the localization in azimuth. Localization in elevation relies on monaural spectral shapes (Musicant et al., 1990, Tollin and Koka, 2009). All those cues are contained in head-related transfer functions (HRTFs), determined by acoustic filtering properties of the head and pinnae (Gardner and Gardner, 1973).

* Corresponding author.
E-mail address: Jakeh.orr@slu.edu (J. Orr).

https://doi.org/10.1016/j.heares.2023.108884
0378-5955/© 20XX

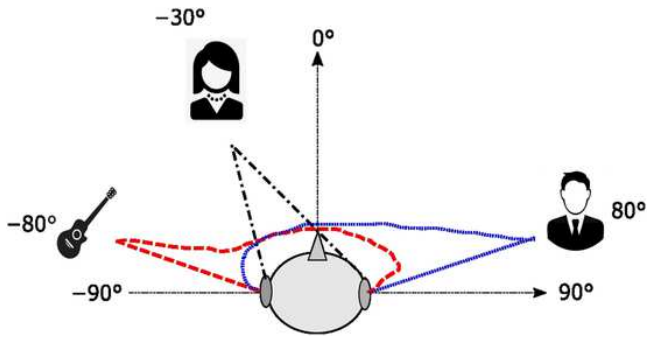

Fig. 1. An example of a scenario with three active sound sources in azimuth: a female speaker at –30°, a guitar at –80°, and a male speaker at 80°. The signals are linearly summed up at each ear of the human with certain amplitudes and timing.

Sound localization based only on ITDs and ILDs suffer from front-back confusions and room reverberations (Halupka et al., 2005, Macdonald, 2008, Wang et al., 2016). Recently, localization algorithms using the entire HRTFs have been developed to predict the locations of multiple sound sources using two (Wang et al., 2016, Rothbucher et al., 2012, Keyrouz, 2017) or four microphones (Keyrouz, 2014, Keyrouz, 2015). The advantages of performing localizations using the entire HRTFs include the ability to localize sources in elevation, higher resilience to room reverberations, and reduced front/back confusions.

Keyrouz and colleagues performed a localization study using binaural microphones (Keyrouz, 2017, Keyrouz, 2008), in which sound locations were determined by forming time-frequency clusters and matching ITDs and ILDs in a HRTF database over a large frequency range. They achieved an averaged angular error around 10° in both azimuth and elevation for multiple concurrent sources. However, their algorithm required information processing over a large range of sound frequencies, which could be time consuming and unsuitable for real-time hearing devices.

Here we present an innovative model as a tool to evaluate the algorithm developed by Keyrouz and colleagues in complex situations that can arise in artificial localization. The key is a normalization procedure that confines all the time-frequency points inside the unit circle. However, it is unclear how those data points behave with various sound locations and frequencies. To thoroughly understand the model behavior, series of simulations were performed in the present study to demonstrate the clustering behavior of the algorithm and to rapidly predict horizontal locations of multiple sound sources based on the spiral pattern generated by our model.

We found that, at a given frequency, the model predicts a spiral patten as the location of a sound moves in the azimuth from left to right. We also found that, with increased model frequency, the spiral pattern evolves from less than one turn into multiple turns. At any fixed frequency, though, the spiral pattern is quite robust over different sound elevations and sound profiles.

In addition, using a room-impulse-response database, the effect of room reverberations was demonstrated. The low-frequency model prediction was accurate under moderate reverberations. As the model frequency increased, errors can occur within adjacent turns. The model failed under strong reverberations, which is reasonable because each sound source can generate multiple echo locations in a highly reverberant environment.

Note that only one model frequency is needed to make a prediction of the horizontal location. This frequency can be arbitrary, as long as enough energy is present in that frequency band. However, when possible, frequencies in the range of 300 to 600 Hz are preferred. Frequencies lower than this range will create location centers that are too packed in the feature space. In contrast, frequencies higher than this range will require a spiral pattern of multiple turns, which is especially problematic for localizations under significant reverberations.

2. Methods

2.1. Sound stimuli

All simulations and signal processing were performed in MATLAB (MathWorks, MA) at a sampling frequency of 44 kHz. A single broadband noise (0–22 kHz, 33.6 s) was used to create accurate sound-location cluster centers at various model frequencies. The broadband

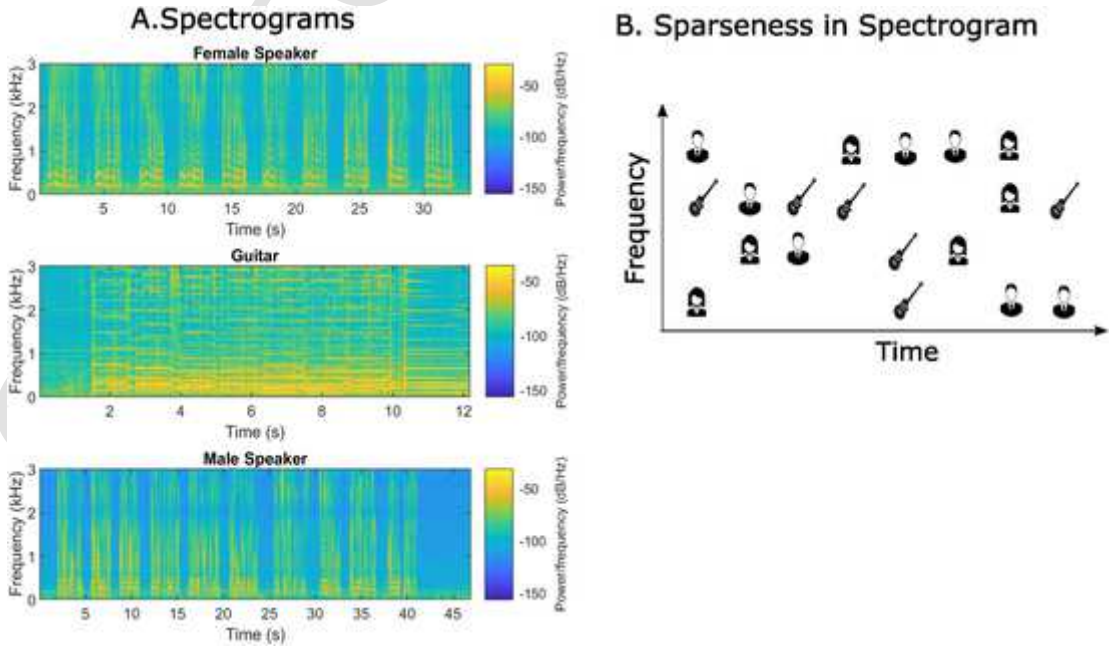

Fig. 2. A, Examples of spectrograms for the three sound stimuli used in the simulations. Each stimulus had a different length (female speech 33.6 s, guitar sound 12.2 s and male speech 52.5 seconds). During simulations, sound files were truncated to make the lengths equal to the shortest sound signal. B, illustration of the sparseness property for natural sound. At a given time frame, there is usually no more than one sound source.

noise is preferred when creating the model stereotypes because, unlike speech sound that has sporadic energy at high frequencies, the broadband noise can guarantee sufficient energy at any frequency. It should be noted that the noise was only used during creating the cluster centers. It is impossible to predict the location of the noise when it occurs concurrently with any other sound because it violates the sparseness property. In addition, a long-duration noise was used in creating the model, but the algorithm works even when the duration was shortened to 150 ms (discussed later).

Four types of sound stimuli (Fig. 1) were then used as "test sound" in the simulation process. First, a sound wave containing a guitar playing, an English speech file by a male speaker, and English speech by a female speaker. The female speech was 33.6 s, male speech 52.5 s, both are reading random lines of a script, i.e., male: "paint the sockets in the wall dull green…" female: "glue the sheet to the dark blue background…"; guitar was 12.2 s and included soft intermediate strumming of an acoustic model guitar. The female speech sound was mostly used during the simulation. To examine how well the result generalizes to other speech sound, we also randomly selected five sentences recorded by a previous study (Calandruccio and Smiljanic, 2012), and concatenated the five sentences into one long speech while running the model. A 5-ms cosine ramp was applied to the onset and offset of the signals.

When the sound-duration effect was examined, sound stimuli were shortened by taking the beginning of the signal, truncating to match concurrent stimuli, and adding the 5-ms ramp. When multiple concurrent sounds were tested, the durations of the longer sounds were shortened to match the shortest sound. Specifically, when the simulation only included male and female speech, male speech was cut short in MATLAB by taking a sample of the first 33.6 s, to make it the same length as the female speech. Whenever the simulation had guitar music in it, the others were cut short in MATLAB to be equal to the guitar music's length, which is the shortest length among the three (12.2 s) in the simulation since there was no inner-ear nonlinearity involved.

Fig. 2A shows the spectrograms of the three stimuli up to 3000 Hz. The female speech had a higher pitch (i.e., the fundamental frequency represented by the first horizontal stripe) than the male speech (Gelfer and Mikos, 2005). The harmonics are multiples of the fundamental frequencies. In comparison, the music has more dispersed frequency components.

2.2. Simulation of sound locations to create the model

As mentioned above, broadband noise violates the sparseness property and cannot be separated from other concurrent sound. However, it has uniform frequency distribution and thus is ideal in creating the localization model. In contrast, a specific piece of speech or music may contain "holes" in their frequency spectrum.

When creating the broad-band location model, HRTFs are needed to simulate the anechoic condition. Because the measurement of HRTFs is cumbersome, an algorithm that works with a standard HRTF database, rather than being fitted to individual humans, will be desirable for hearing devices. The CIPIC database (Algazi et al., 2001) was used in this study. The database contains HRTFs from –80° to +80° in azimuth, with positive values being the right side and negatives on the left. Angular separations were approximately 5°. The distance of each sound source was fixed as 1 m, according to the distance of the CIPIC database.

The CIPIC database contained a total of 45 sets of HRTFs obtained from 45 subjects. We randomly selected Subject 21 to create the model; later this model will be used to predict sound locations for other subjects. The horizontal angle of interest was simulated by filtering a mono sound with the HRTF pair that corresponded to the desired horizontal angle. The simulated angle in elevation was kept at 0° for all the plots except Fig. 7, in which other elevations were simulated.

2.3. Experimental paradigm

In Experiment I, the listening environment was assumed to be an anechoic chamber; the effect of room reverberations was ignored. Speech and/or music sound were used as "test sound" while the location centers derived from the broadband model were applied to predict the locations of the test sound.

In Experiment II, reverberations were added to the anechoic signal. Room impulse responses were obtained from the Aachen Impulse Response (AIR) database (Jeub et al., 2009, Jeub, 2019). AIR contains binaural room impulse responses (BRIR) for different listening environments. The BRIRs were derived in a way similar to HRTFs: sound coming from various speakers was recorded using binaural microphones that were placed inside a HMS2 artificial head by HEAD acoustics. The difference was that the room was not anechoic. In other words, BRIRs contain both HRTFs and the room impulse responses. By filtering a mono sound with a pair of BRIRs from the AIR database that corresponded to our desired reverberation level and horizontal location, the effect of reverberation was observed.

Here, two reverberant environments were simulated.

Stairway Hall is a listening environment with moderate reverberations ($RT_{60} = 0.83\ s$). The reverberation time, RT_{60}, is defined as the time period after the termination of a sound when the sound reaches a 60-dB attenuation. The longer this time, the stronger the reverberations. An acceptable RT_{60} time is dependent on the room type and what it will be used for, as a music hall would not want to have zero reverberation, while a lower amount of reverb would be desired for a classroom setting. In general, an RT_{60} of <1-1.5 s is ideal for a regular speaking environment. The database for this room includes BRIRs from –90° to +90° in azimuth with angular separations of 15°. The source distance was also 1 m.

Next, we simulated the reverberant listening environment, *Aula Carolina*, a former church with a ground area of 570 m^2 and a high ceiling. This room has strong reverberations with $RT_{60} = 5.16\ s$. The speaker was at a distance of 3m. The BRIRs ranged from -90° to +90° with angular separations of 45°.

Again, the female speech was used as "test sound" while the location centers derived from the broadband model were applied to predict the locations of the test sound under room reverberations.

2.4. Phase normalization and the clustering algorithm

The localization algorithm developed by Keyrouz and colleagues (Keyrouz, 2017) is based on a feature space related to ITDs and ILDs at various sound frequencies. First, the recorded sound signals from binaural microphones are transferred into the time-frequency domain by means of Short-Time Fourier Transform (STFT). During the process of STFT, the long signal is divided into overlapping frames, and the Fourier transform of each frame is computed. Regarding the integration window size for the STFT, the initial study (Keyrouz, 2017, Keyrouz, 2008) used less than 40 ms. As will be shown later, 40 ms generated relatively large scattering, and we found that 80 ms can yield an almost ideal result. Therefore, we used 80 ms as the integration window throughout the study, except for Fig. 5A where we demonstrated the above-mentioned scattering with the 40 ms window.

Performing the STFT results in a series of complex numbers. The magnitude and phase can then be subsequently derived from those complex numbers as $|X|$ and φ independently for the left and right channels. The localization algorithm then divides each time frame into frequency bands to perform localization at each frequency band separately. The key is a normalization procedure applied to the right-microphone recording,

$$X_R(f,t) = \frac{X_R(f,t)}{\sqrt{|X_L(f,t)|^2 + |X_R(f,t)|^2}} e^{-i\varphi_L} \quad \text{(Eq. 1)}$$

Afterwards, the phase becomes the difference between the right- and left-ear phases ($\varphi_R - \varphi_L$), which is usually called the interaural phase difference (IPD) and is uniquely related to the ITD at a certain frequency. Here, φ_L is the phase of the left recording. The real and imaginary parts of the normalized X_R are typically plotted against each other for each frequency band.

Examining the distance of the moving cluster's center to the fixed origin, it depends on which ear is louder. When the left ear is louder (source closer to the left ear), the cluster is closer to the origin. Fig. 3 is a schematic plot of how the time and level cues work in determining the feature space. For each sound source, if the right ear leads the left ear, the normalized phase would be positive (counterclockwise in Fig. 3A). This phase cue is most robust at low frequencies. If the right ear is louder, $|X_R|$ would approach 1 (Fig. 3B); if the left ear is louder, $|X_R|$ would approach 0. Equal amplitudes (ILD = 0) lead to $|X_R| = \frac{1}{\sqrt{2}} = 0.7$ (Fig. 3B, *green dotted circle*). This level cue is most robust at high frequencies. The behaviors of the points would change according to the low/high frequencies and whether ITD or ILD dominates, however, we apply the clustering algorithm in the same manner.

For the scenario shown in Fig. 1 with three active sound sources, the three corresponding clusters were shown in Fig. 4 over a large range of frequencies. Each point (Fig. 4, dots) corresponds to a scaled value of the STFT at a certain time and frequency. The importance of sparseness for concurrent sound localization lies in the fact that each data point in the feature space represents only one sound source in a short time window (Fig. 2B). All the points are inside the unit circle in the feature plane due to the normalization procedure implanted (Eq. (1)). The center of each cluster (Fig. 4, large circles) supposedly corresponds to the scaled value of the HRTF for a particular sound-source location. This is because each HRTF is unique to the location of the sound source, and, for a stationary source, the transfer function of the sound source to ear remains the same over time frames. The number of individual centers corresponds to the number of active sound sources in that frequency band.

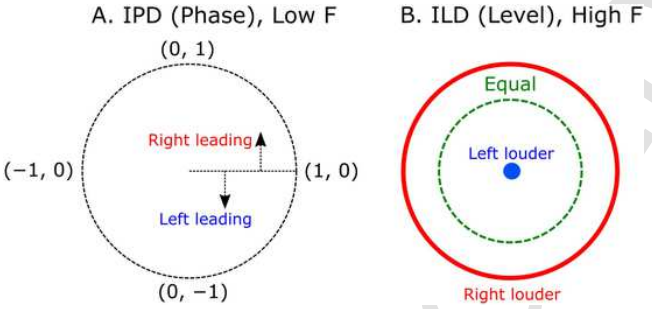

Fig. 3. Normalized feature space showing the ITD (3A, left) and ILD (3B, right) cues at low and high frequencies.

However, there is no easy way to identify which cluster center corresponds to which sound source, e.g., the guitar at −80°, the female speaker at −30°, or the male speaker at 80°. Previous studies did not systematically examine how those cluster centers vary with sound locations. This is precisely the reason why we performed this study—to create a model at various frequencies that allows us to easily identify the sound-source location(s).

Next, a k-means approach was performed using MATLAB's standard "k-means" function with five "Replicates". This function automatically searches in the normalized feature space and identifies the center location of each cluster for one to three clusters, depending on the simulation condition.

2.5. Horizontal-location classification

As mentioned above, there were 45 HRTF sets in the CIPIC database, and the model was created by arbitrarily choosing Subject 21's HRTF. Under the anechoic condition, the model created using broadband noise was validated using the first seven subjects' HRTFs in the same database for single or multiple speech sound. Localization was performed over the same range of angles as was used in the model creation, e.g., –80°, –65°, etc. and at multiple frequencies. Here, each newly generated test cluster (location unknown) for a novel subject was assigned to the nearest cluster center (location known) in the model, and thus a horizontal angular location can be determined. Here, "nearest" means that the smallest Euclidean distance of this test center to all the model clusters was selected as the target. For each test data input, a "localization error" can be computed by taking the absolute difference between the true angular location and the classified angular location using the model. In other words, the error is the systematic error between the pre-

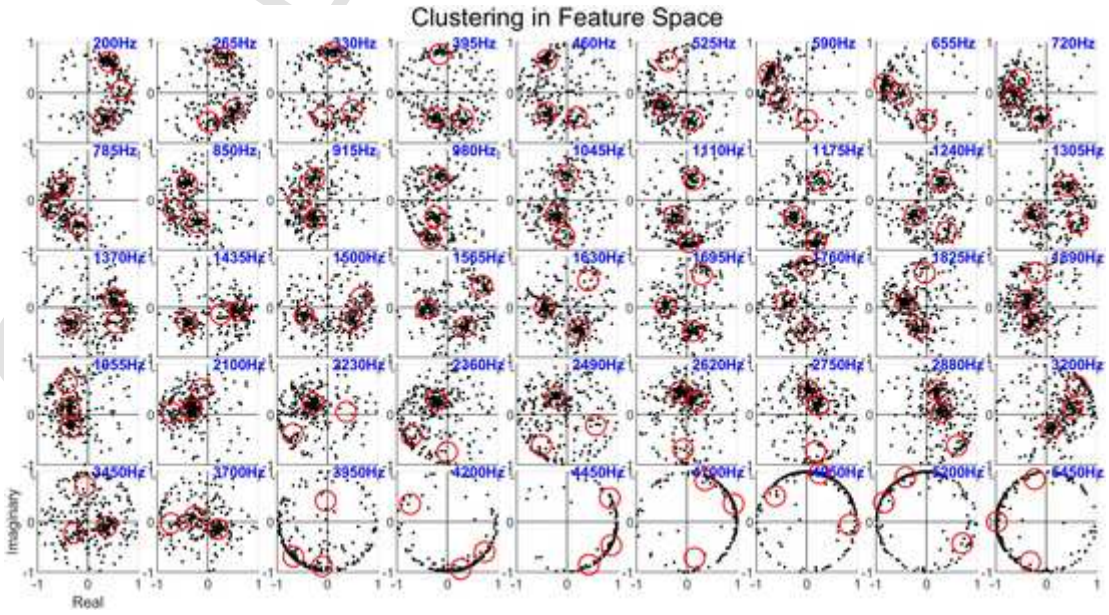

Fig. 4. Three clusters in the feature space for the three sound sources shown in Fig. 1 covering frequencies of 200 to 5450 Hz. Each cluster center (i.e., a red circle) supposedly corresponds to one of the source locations, –30°, –80°, or 80°. The. The elevation was fixed at 0°.

dicted sound location and the true location. The true location determines the pair of HRTFs used to filter the sound stimuli.

Next, the localization error was measured with a single speech sound under reverberant conditions. We did not perform multi-sound localization under reverberations because, as will be shown later, even a single source was difficult to classify.

3. Results

3.1. The broadband model

To create the basic model and study how a cluster's position changes according to the sound location in azimuth, an anechoic listening environment for various locations in the frontal hemisphere was first established. As stated earlier, a single broadband noise filtered with a set of HRTFs was used to establish the location centers covering the range of –80° to 80°.

Fig. 5 shows the model behaviors for four representative frequencies, 300, 600, 1500, and 3000 Hz. As mentioned earlier, we used 80 ms as the integration window for STFT since it created less scattering in the model. In Fig. 5A (right), we also plotted the same model when a shorter time window (40 ms) was used to demonstrate the effect of window size. Regardless of the window size, it can be seen that, as the horizontal location of the broadband noise moved from left to right, the clusters traveled in the feature space in a counterclockwise manner. For the lowest frequency, 300 Hz, the spiral was roughly a half turn (Fig. 5A). As the model frequency further increased, the spiral reached one full turn and eventually multiple turns (Fig. 5, B–D).

The spiral pattern was codetermined by both ITD and ILD cues. Briefly, the distance of each cluster center to the fixed origin (0, 0) depends on which ear is louder. When the left ear is louder (source closer to the left ear), the cluster is closer to the origin as seen in Fig. 3B, where $|X_R|$ approaches 0. When the right ear is louder, the cluster is closer to the unit circle, and $|X_R|$ would approach 1 (Fig. 3B). The ITD determines the counterclockwise turning property. Overall, the results showed that the cluster's center position follows a spiral pattern that can help predict the azimuth of the active sound source with only one frequency.

Here, we used broadband noise to create the model because it has a uniform energy distribution over frequencies. It is thus important to verify that the observed spiral pattern did not show sensitivity to the type of sound stimulus. Fig. 6 shows the raw data points (dots, one color for each location) after the normalization procedure for the four stimuli tested in this study (i.e., A, broadband noise; B, female speech; C, male speech; D, music). The shell-like spiral shape was almost identical when the frequency was fixed at an arbitrary value, 655 Hz, but the raw data points were more scattered in the speech and music plots (B–D) than in the broadband noise plot (A). Here, the large circles in all the plots were derived from the broadband noise data and superimposed on the other three conditions. The elevation was fixed at 0°.

Note that the cluster centers are not the geometrical means of each location, which would have been affected by certain outliers. Instead, each center was identified by the k-means approach as where the *density* of dots was highest.

3.2. Experiment I: dual sources and different elevations under anechoic conditions

The above basic model was derived with a single broadband-noise source and fixed elevation of 0°. Next, two active sound sources were simulated, and the elevation was varied. Because the broadband noise violates the sparseness property, speech sounds were used here. In Fig. 7, one source (the female speech) was kept stationary at –80° in azimuth (i.e., to the left of the virtual listener) and the second source (male speaker) traveled from one stationary location to the next, covering –80° to +80° (i.e., toward the right side of the virtual listener) in in-

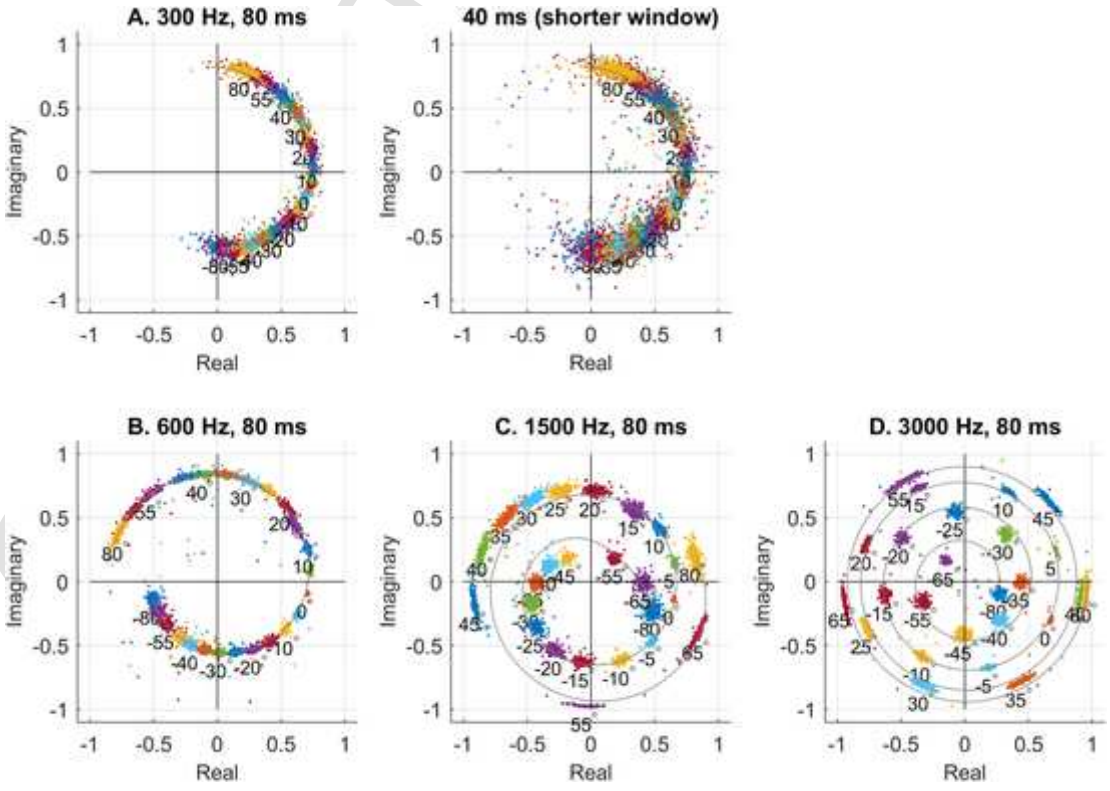

Fig. 5. The broadband-noise model that displays changes in the spiral patterns with increasing model frequencies. At very low frequencies (A), clusters are formed in a tight pattern with minimal spirality (half turn). As the frequency increases (B–D), clusters begin to spread further apart, and spirality evolves into multi-turns. In A, two analysis-window sizes are explored, 80 ms vs. 40 ms. Since the 40 ms generates larger scatters, we will use 80 ms for the rest of the study.

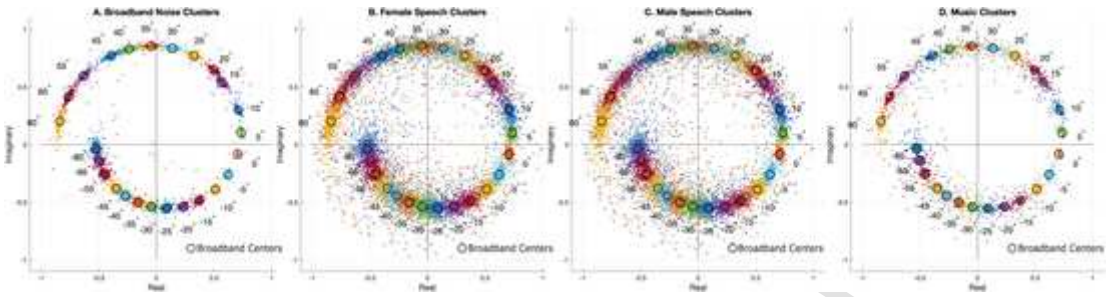

Fig. 6. Model simulations of cluster centers at an arbitrary frequency (655 Hz) for different sound profiles. The broadband cluster centers are superimposed on each of the four conditions. In general, the cluster locations in the complex plane were not notably sensitive to the type of stimulus. This provides evidence that the single-source broadband model can be used to predict the location(s) of single or concurrent sounds with various profiles. The elevation was fixed at 0° in the simulation.

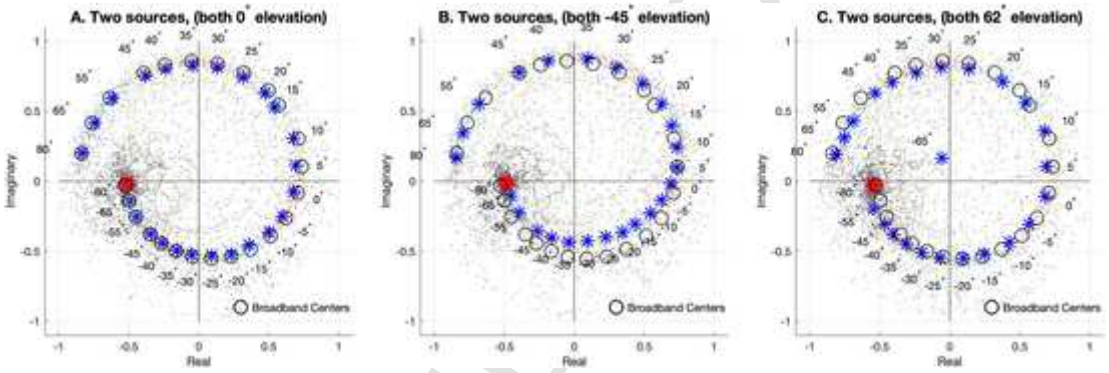

Fig. 7. Experiment I: localizing two sound sources at 655 Hz under anechoic conditions. One source (female speech) was kept at –80° (red symbol). The other source (male speech) moved from –80° to +80° in azimuth. The cluster moved on a pattern that ultimately formed a shell-like shape. Note there were multiple red asterisks that were overlapping with one another.

crements of 15° or less. At each location, localization was performed for an arbitrary frequency of 655 Hz, and the position of the moving cluster center relative to the static source was observed in the complex plane (Fig. 7). Here, the two source locations were simultaneously identified using the k-means approach.

In Fig. 7A, both sources were kept within the 0° elevation plane. The red asterisk indicates the cluster location for the direct source kept at –80°. Note that there are multiple red asterisks overlapping on one another. The blue asterisks are the direct cluster locations for the moving source, which formed a counterclockwise spiral pattern. The black circles are single-source locations derived from the broadband-noise model presented above. Given that moving-source centers more or less matched the single-source centers, we conclude that the k-means approach worked successfully extracting both source locations concurrently.

Next, we repeated the simulation at two more elevations. In Fig. 7B, both sound sources were lowered to –45°, which was the lowest elevation in the front hemifield of the CIPIC database. The black circles are again superimposed single-source model centers at 0° elevation. Although there were disparities between the two elevation results (Fig. 7, A and B), the model structure generally held. When the two sources were moved to 62° elevation, observations were similar, except that for one source, –65°, the k-means approach failed to identify the moving-source location. This is due to the fact that the fixed-source location (–80°) is not much different from the moving-source location (–65°) when both sources are elevated by 62°; that is, the two locations reside closely on a small circle above the head of the virtual listener. The HRTFs measured for those two locations are presumably quite similar.

In Fig. 6 we showed that the spiral patten did not vary much with the sound profile. If one prefers constructing the model using speech sound, it should work equally well at relatively low frequencies where enough sound energy is present in the speech sound. Fig. 8 shows an example of predicting two source locations using a speech model at a single model frequency. Again, we used the arbitrary frequency 655 Hz as an example. Fig. 8A shows the mixed-sound situation when the female-speech source was located at –45° and the guitar was at –80°. If we compare the raw data points (Fig. 8A) to the same female-speaker model replotted from above (Fig. 6B), we can identify which cluster belonged to the source at –45° and which belonged to –80°. The large circles are the cluster centers identified by the k-means from the mixed sound superimposed on the female-speech model plot (Fig. 8B), indicating an accurate match of the locations.

3.3. Experiment I: localization errors and the duration effect

In the above model construction, broadband noise, and a random subject's HRTFs from the CIPIC database were used to generate the cluster centers. The broadband model in Fig. 6A that was used for the rest of the study had a sound duration of 33.6s. Whereas to evaluate the model performance under anechoic conditions, the speech and music serving as test sounds also had durations more than 10-s long. Here, we first applied the model to make predictions for seven other subjects in the CIPIC database. Meanwhile, we examined the effect of test-sound duration on the model accuracy.

We found that the shortest duration that can produce the cluster behavior sufficient for our algorithm to make a reasonable prediction was around 150 ms. Durations shorter than that would likely generate a significant error even under “easy” conditions, such as a single sound under the anechoic condition. Fig. 9A applied the model constructed with a random subject (Subject 21 in the CIPIC database) to predict sound locations using HRTFs recorded with seven novel subjects (also from the CIPIC database) for durations of 10 s (Fig. 9A, top row) vs. 150 ms (Fig.

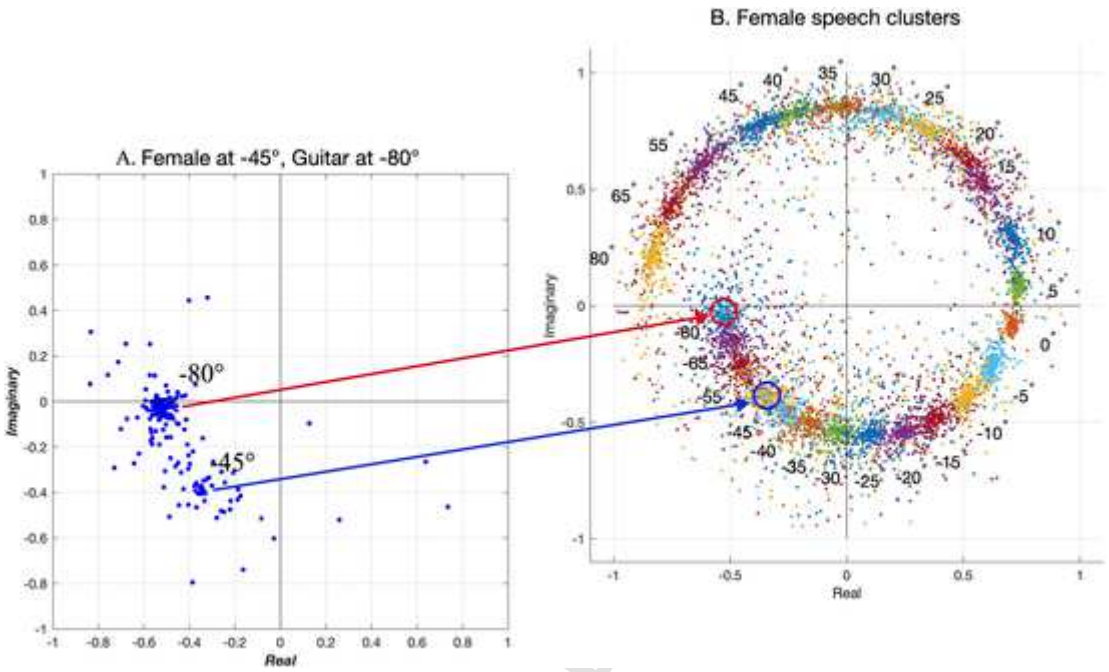

Fig. 8. Comparisons between raw data (A) and the model clusters (B). *A*, simulation with the female speaker fixed at –45° degree and the guitar at –80°. *B*, when the broadband model was re-run with the female speech.

9A, bottom row). The x-axis was the true angular location associated with a particular HRTF. The y-axis was the classified location using the broadband model at various frequencies (Fig. 5).

Generally speaking, under the anechoic condition, the model at low frequencies (i.e., 300 and 600 Hz) performed better than at high frequencies (1500 and 3000 Hz). Recall that the spiral pattern began to evolve into multi-turns above 1 kHz (Fig. 5). This is not a problem for the broadband noise, as the clusters were quite focused for broadband noise. However, speech sound tended to have more dispersed clusters, as illustrated in Fig. 6 (B and C), even though the cluster centers more or less matched the centers derived from the broadband model. Consequently, as the model frequency and the number of spiral turns increased, more localization errors appeared due to misclassifications across the turns (Fig. 9, right two columns). In fact, even at 600 Hz, some subjects already showed misclassifications between the 80° and –80° locations (Fig. 9, second column). In summary, under the anechoic condition, low model frequencies generated better performance. The predicted angles were closest to the true values at the model frequency of 300 Hz (Fig. 9A) with an average error of 4.5° and RMS error of 6.4°.

Meanwhile, there was no duration effect as long as the signal duration was no less than 150 ms. No statistical difference (*t* test) was found between the classification results for the two durations, 10 s vs. 150 ms (Fig. 9A, upper and lower rows). In other words, the "near-real time" of this model when applied to real-time hearing devices would be around 150 ms.

Up to this point, the model has been established using HRTFs from a random human subject (Subject 21) in the CIPIC database, and the result was mainly obtained with the female speech sound. In Fig. 9A, the model created with this subject was used to predict sound locations for eight new human subjects. To examine whether the choice of the model subject affects the model performance, another subject (Subject 48) was randomly chosen from the database to establish the model and the localization was repeated (Fig. 9B). Meanwhile, five speech sentences were randomly selected from recordings in a previous study (Calandruccio and Smiljanic, 2012) and concatenated into one long speech sound. Despite the two major changes, the model performance was qualitatively similar (Fig. 9B). Therefore, we did not consider it a problem when constructing the model with HRTFs from one subject, and we believe our simulation results can generalize well with daily speech sound. All results from experiment I are summarized in Table 1 below.

To evaluate the model performance, we also repeated the experiment (Fig. 9A) using a Generalized Cross-Correlation Phase Transform (GCC-PHAT) approach (Knapp and Carter, 1976). The same set of human HRTFs used in Fig. 9 was applied to create the virtual sound locations. The standard MATLAB function, gccphat, was used to compute the time delay between the simulated left- and right-ear sound. Next, the horizontal angle was determined (Ollivier et al., 2019) given that the average distance between the two ears was 14.5 cm and the speaker rack had a radius of 1 m in the CIPIC database.

Using a 1-s broadband noise as the single-source stimulus, the performance of the GCC-PHAT algorithm was satisfactory (Fig. 10A), which proves that the method generally works. However, when using the same 0.15-s speech sound as used in Fig. 9A, the GCC-PHAT performance was notably worse. It either predicted a location similar to the broadband condition (Fig. 10A), or a location at the front center regardless of the true location. For those error trials, the delay identified by the GCC-PHAT was always 0; that is, it failed to yield a non-zero delay. The performance was certainly worse than our low-frequency models (Fig. 9).

3.4. Experiment II: the reverberation condition

The above simulation shows predictions for ideal localizations in an anechoic chamber. In real life, reverberations often create echoes that affect both the timing and level of the received sound. Here, we simulated two types of reverberations to examine the effect on localization accuracy using the above-mentioned spiral model. After filtering a mono sound with BRIRs chosen from the AIR database, the resulting sound contains both room reverberation and binaural cues that are the result of filtering by the listener's head (same cues that are contained in HRTFs).

Fig. 11 shows an example of the female sound with different virtual locations using the model frequency 655 Hz under the reverberation condition, *Stairway*. The simulation generated moderate reverberation ($RT_{60} = 0.83\ s$). Here we plotted four locations, –45°, –30°,

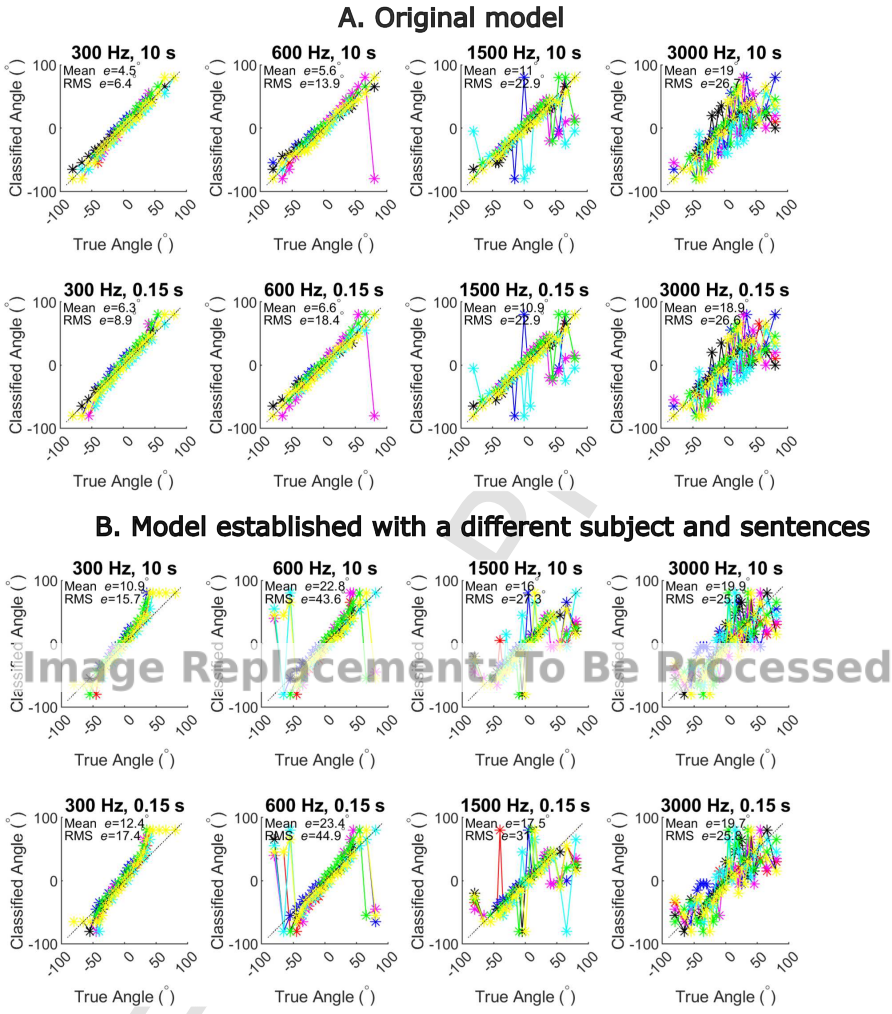

Fig. 9. Experiment I: sound-localization performance in the anechoic condition for different sound durations and model frequencies. A. the model was created using HRTFs from a random human subject (Subject 21), just as the rest of the study. The first seven subjects' HRTFs were selected from the CIPIC database as the test subjects. For each condition, the grand average error, Mean *e*, and the RMS error were shown in the figure. The sound stimulation was the female long speech. B, the model was recreated with HRTFs from Subject 48 in the CIPIC database. The sound stimulation was five random sentences recorded by an external study (Calandruccio and Smiljanic, 2012).

Table 1
Experiment I Results.

Experiment I: A. Original Model			
Frequency (Hz)	Duration (s)	Mean (e)	RMS (e)
300	10	4.5°	6.4°
600	10	5.6°	13.9°
1500	10	11.0°	22.9°
3000	10	19°	26.7°
300	0.15	6.3°	8.9°
600	0.15	6.6°	18.4°
1500	0.15	10.9°	22.9°
3000	0.15	18.9°	26.6°
Continue2A Model established with a different subject and sentences			
Frequency (Hz)	Duration (s)	Mean (e)	RMS (e)
300	10	10.9°	15.7°
600	10	22.8°	43.6°
1500	10	16.0°	27.3°
3000	10	19.9°	25.8°
300	0.15	12.4°	17.4°
600	0.15	23.4°	44.9°
1500	0.15	17.5°	31.0°
3000	0.15	19.7°	25.8°

15° and 60° (there are a total of 13 locations in this reverberation database to choose from) superimposed with the female-speech model. That is, both the model and the reverberant simulation were generated with the same speech so that it is easier to compare the amount of data-point dispersion in the anechoic condition (the model) with the reverberant condition. The position of the cluster under reverberation is shown with a green asterisk. It can be seen that even though the reverberant room type was used, the cluster was more or less close to the expected cluster from the model. The exact errors for different source locations and different frequencies will be presented in Fig. 13.

Fig. 12 shows the result for angles –90°, –45°, 0° and 45° for *Aula Carolina*. This listening environment has strong reverberations due to a high ceiling and architecture of the church. Additionally, the speaker was very far from the microphones (9.8 ft). In other words, there were essentially multiple sound sources (i.e., the origin and loud echoes) present in the listening condition, while the model was trying to come up with a *single* location using the k-means approach. The predicted single source was not close to the true source for those peripheral angles (Fig. 12).

Similar to the anechoic condition (Fig. 9), to quantify the localization performance under reverberations, we classified the locations based on clusters elicited by reverberant sound according to the anechoic broadband spiral model. Fig. 13 shows the classified horizontal locations as a function of true angles used in the simulation for both reverberation conditions at multiple model frequencies.

For the *Stairway* simulation, the reverberation was moderate with $RT_{60} = 0.83\ s$. The predicted angles were closest to the true values at the model frequency of 600 Hz (Fig. 12B), with an averaged mean error of 15.4° and RMS error of 21.5°. At high frequencies (i.e., 1500 and 3000 Hz), the model failed to predict the locations. As shown in Figs. 10 and 11, reverberations always reduced the vector length, i.e., making the cluster closer to the center of the feature space, (0, 0). Therefore, where there were multi-turns at high frequencies, the shrinking effect would lead to large errors.

It is also interesting to observe that the model performed worse at the lowest frequency, 300 Hz, than at 600 Hz. This was due to the fact that, when the frequency was too low, the cluster centers were too packed together (Fig. 5A). Therefore, although lower frequencies worked better than higher frequencies using our model, it should not be perceived as the lower, the better.

The model completely failed to predict the locations under the *Aula Carolina* condition, which was strongly reverberant with $RT_{60} = 5.16s$ (Fig. 11B). We only simulated five horizontal locations, because only those five were available in the reverberation database. Here the mean classification error was between 39.6° and 61.2° and RMS errors ranged from 63.3° to 78.8°, but more importantly, all angles were classified as very negative/leftmost angles. All results for experiment II are summarized in Table 2 below.

4. Discussion

4.1. Novelties and applications of the proposed model

In many normal situations and everyday interactions, we find ourselves in environments where sounds are coming from multiple sources and locations. Although normal human listeners can localize more than one sound source concurrently (Zhong and Yost, 2017, Keller and Takahashi, 2005), this is not the case for the hearing impaired. Furthermore, standard hearing devices and aids, including modern cochlear implants, do not restore a normal level of sound-source localization for hearing-impaired listeners (Loiselle et al., 2016). Sound-processing algorithms that can automatically decode sound locations and subsequently segregate the sound into different streams according to the locations would be highly desirable for hearing aids.

In this study, the focus is on the localization of sound sources in the azimuth. Segregation of sounds would require a more involved process where each frequency in the listening space is examined. For example,

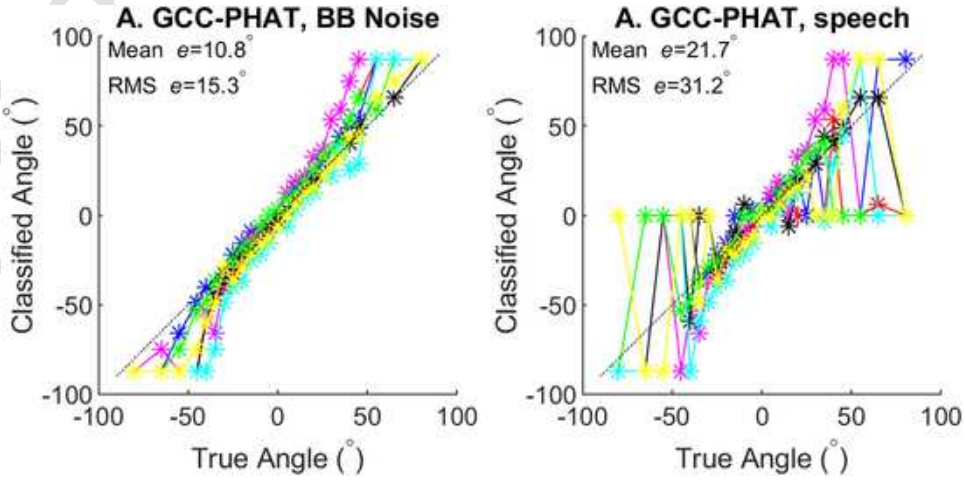

Fig. 10. Simulation results using a standard GCC-PHAT model. Results were obtained with the same set of human HRTFs used in Fig. 9. The single-source sound was either a 1-s broadband noise (*A*) or the female speech (*B*) same as what was used in Fig. 9A.

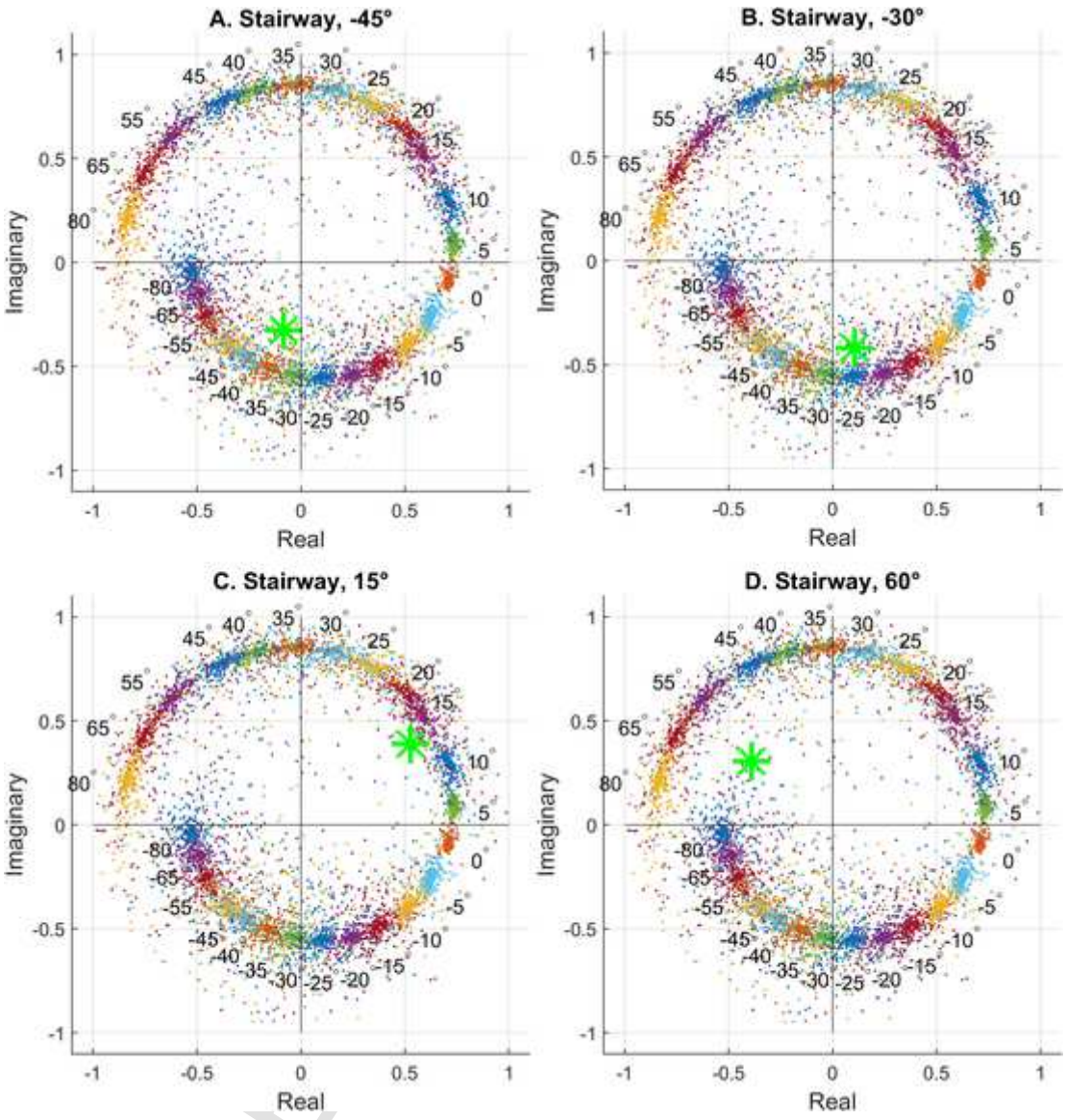

Fig. 11. Experiment II: simulation result with the female speech under room reverberations, *Stairway Hall*, as a green dot superimposed on the previously derived model at 655 Hz. Four horizontal locations were selected for the demonstration. Moderate reverberations were added by setting RT60 = 0.83 seconds.

clustering needs to be performed for each frequency in Fig. 4, after which a "binary mask" can be constructed to remove the unwanted source. The end result would be a sound localization map for the user of the hearing aid, and the user can select sounds they wish to remove from specific locations in their environment. While our simplified model does not include the segregation and removal of sounds, it does allow localization of the sound. This is particularly useful for hearing impaired listeners who are crossing a street and need to know the direction of a siren contained on a fast-approaching emergency vehicle, as one example.

Previous studies have investigated other methods of sound source localization, using more than two microphones. With four microphones (Keyrouz, 2015, Keyrouz, 2008), one inside and one outside the ear canal on either side, monaural HRTFs can be derived. Source locations are determined by finding the best matching HRTFs. Keyrouz and colleagues also performed a localization study using binaural microphones (Keyrouz, 2017, Keyrouz, 2008), in which locations were determined by matching ITDs and ILDs in a HRTF database over a large frequency range. Using an unsupervised clustering algorithm, such as the multiple self-splitting and merging algorithm (Liu, 2007), the center of each cluster, and therefore the estimated value of scaled HRTF in the corresponding frequency band, can be identified. Reconstructed HRTFs using the centers across *all the frequencies* can then be used to search for the best-matched HRTFs in the CIPIC database to determine the sound locations.

Additional previous studies have also examined binaural sound source localization in the field of robotics, where localization is carried out over a broad frequency range. (Lyon, 1984) used a dynamic spectral mask based on energy over very short intervals in binaural cross-correlation. They were able to achieve a rudimentary spectral separation of small samples but were limited by computational power. Baumann et al. (2015) focuses on improving the spatial resolution of the auditory scene analysis by rotating or translating the microphones (ears) of a robot over time using binaural microphones. They rely on an algorithm by which instantaneous evidence and prior knowledge are co-registered over successive movements of the microphones (for example, by Recursive Bayesian, Kallman or Particle Filtering approaches). This approach may be beneficial in improving our model in the future, providing it does not require a significant increase in computational power. Finally, (Raspaud et al., 2009) propose a binaural source local-

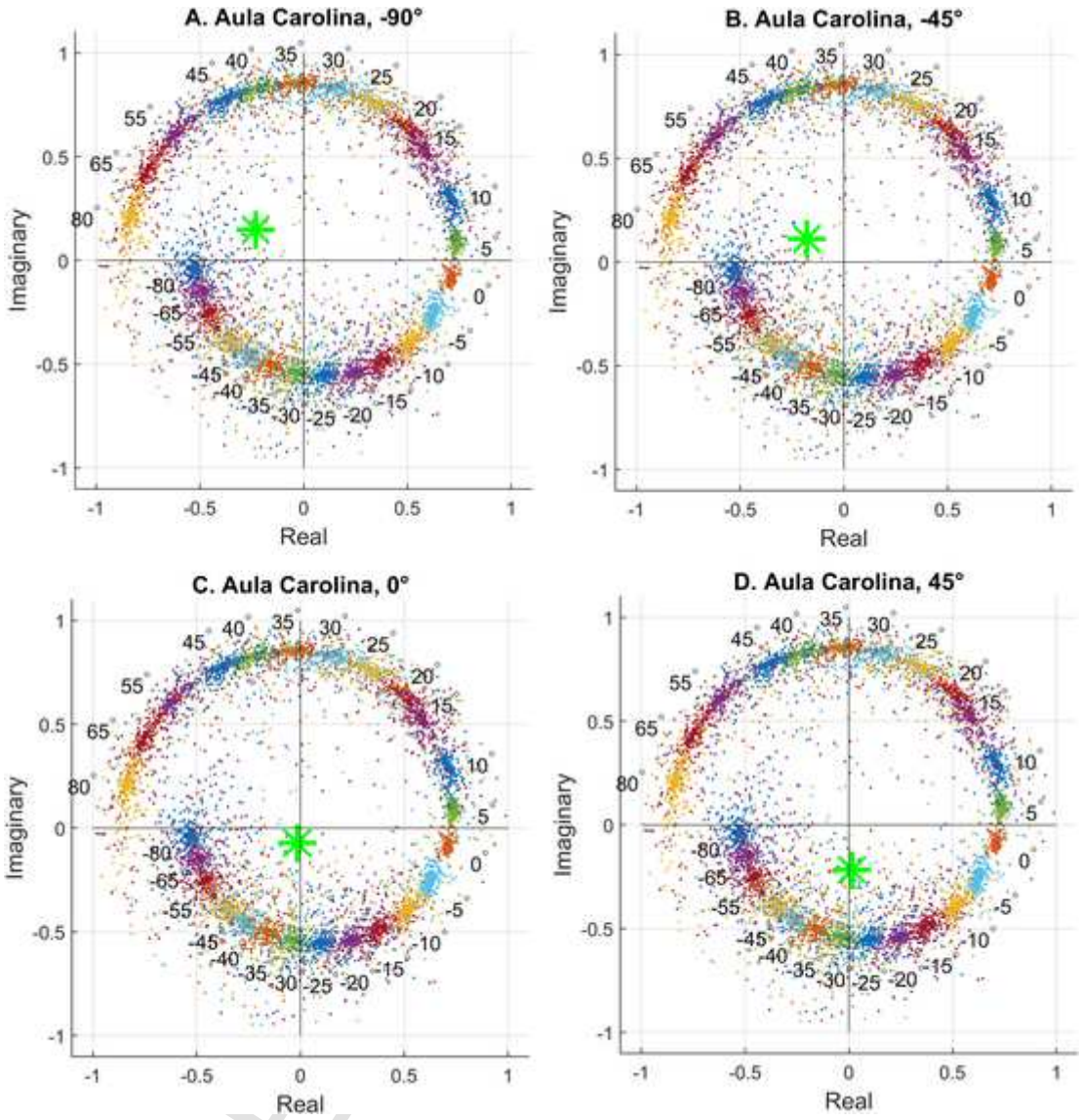

Fig. 12. Experiment II. Same as Fig. 11, except that the reverberation type was *Aula Carolina* and the reverberations were strong.

ization method based on ITDs and ILDs. The two cues, computed from a two-channel time-frequency representation, are combined in order to estimate the azimuth of sources in binaural recordings. The maximal average error for their algorithm ranged from 3.35 to 15.24 degrees, when localizing a single sound source.

Roman et al. (2003) construct an estimate for an ideal binary mask by using a supervised learning approach. The system learns to use both ITD and ILD cues to assess the degree of interference within a particular frequency band at a particular time. Time-frequency regions predicted to be dominated by the target are used to resynthesize the estimated target signal. However, their algorithm focuses only on speech segregation and not source localization.

Kraljević et al. (2020) presents a passive three-dimensional sound source localization (SSL) method that employs a geometric configuration of three soundfield microphones. Two methods for estimating the angle of arrival (AOA) and time difference of arrival (TDOA) are proposed based on Ambisonics A and B format signals. The closed-form solution for sound source location estimation based on two TDOAs and three AOAs is derived. The proposed method is evaluated by simulations and physical experiments in an anechoic chamber. While good localization results were seen in both the 2D and 3D space, this method is not directly applicable to hearing aid devices due to the reliance on 3 microphones, and the correlation between longer time of observance and increased accuracy of localization, with requirement of stationary objects for observance.

Dang et al. (2019) proposes a method to address the multiple sound source localization problem by associating and fusing the direction of arrival (DOA) estimates from multiple microphone arrays. Their solution lies in a multi-dimensional assignment-based data association approach to find the optimal associations of DOA estimates from the same source. First, in the sense of maximum likelihood, the data association problem is formulated by finding the most likely partition of the measurement set into the source-originated and false alarm-originated subsets. Next, by defining the association costs appropriately, the problem of finding the most likely measurement partition is transformed into a generalized multi-dimensional assignment problem which can be solved efficiently by a Lagrangian relaxation algorithm. After the optimal associations of DOA estimates across different arrays are obtained, the locations of sources can be estimated by fusing the same source-originated DOA estimates. While this method is also applicable to the issue we are working to solve, there are a few setbacks that limit this capability for use in hearing aids: the amount of processing required for the data association and optimization algo-

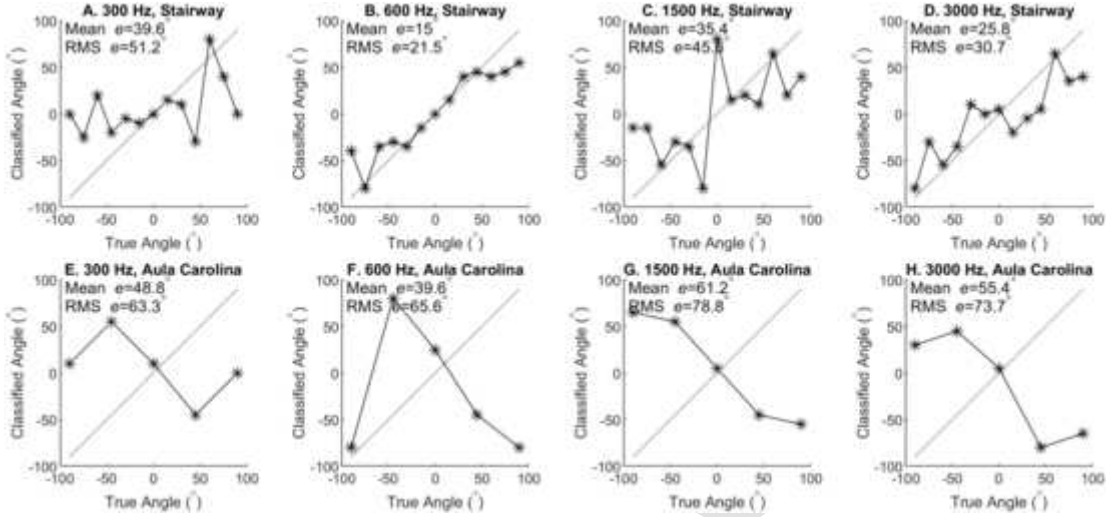

Fig. 13. Experiment II: Classified angles vs. true angles for the two reverberation conditions at various model frequencies. For each frequency, the averaged error, Mean *e*, and the RMS error were shown in the figure legends.

Table 2
Experiment II Results.

Experiment II			
Frequency (Hz)	Reverberation Condition	Mean (e)	RMS (e)
300	Stairway	39.6°	51.2°
600	Stairway	15.0°	21.5°
1500	Stairway	35.4°	45.6°
3000	Stairway	25.8°	30.7°
300	Aula Carolina	48.8°	63.3°
600	Aula Carolina	39.6°	65.6°
1500	Aula Carolina	61.2°	78.8°
3000	Aula Carolina	55.4°	73.7°

rithms would not be suitable for use in hearing aid devices, due to size and computational power limitations.

Additionally, there are studies where sound localization under reverberant conditions are carried out, showing promising results. Grumiaux et al. (2022) investigates several studies that utilize deep learning methods for sound localization under reverberant conditions. Vecchiotti et al. (2019) proposes a novel approach for sound localization, estimating the azimuth of a sound source directly from the raw waveform. Instead of using hand-crafted features commonly employed for binaural sound localization, the authors employ convolutional neural networks (CNNs) to extract specific features from the waveform that are suitable for localization. This method of localization shows promise, however, appears to be too computationally dependent for use in hearing aid applications.

Wang et al. (2019) proposed a method which leverages deep learning techniques to create a time-frequency mask that highlights the regions in the audio signal containing the speaker's information. This mask is applied to the audio spectrogram, which transforms the audio signal into the time-frequency domain. By emphasizing the speaker-related information, the method enhances the localization accuracy, particularly in noisy or reverberant environments. However, there are significant differences between the microphone array-based speaker localization system and hearing aids. Of importance, hearing aids must operate in real-time and have limited processing power and energy constraints that would limit this methods success.

Finally, Ma et al. (2017) proposes a novel approach to robustly localize multiple sound sources in environments with reverberation using a combination of deep neural networks (DNNs) and head movements. As displayed in our current study, localization of sound sources is a challenging task in reverberant spaces due to reflections and delays. To combat this challenge the researchers, leverage binaural audio signals and integrate head movements to enhance the accuracy of sound source localization. By employing DNNs, the system is capable of learning complex spatial features from the binaural inputs. Furthermore, the incorporation of head movements allows the system to adaptively update its source localization estimates as the user moves. However, Hearing aids require reliable, low-latency, and user-friendly solutions tailored to their limited processing power, battery life, and diverse real-world listening scenarios, making the integration of DNNs and head movements introduced in this study less practical for hearing aids.

While the above-mentioned studies are generally applicable to sound localization, the complexity of their algorithms and requirements for storage of large amounts of data indicates a weak point for utilization in simple hearing aid devices; an issue we aim to overcome with a simplified localization model.

In the present study, we present an innovative model as a tool to systematically evaluate the algorithm developed by Keyrouz and colleagues in complex situations, specifically when there are room reverberations. The model is much simpler and faster in determining the horizontal location of multiple sound sources, and the performance is largely invariant to various types of sound (male, female, music, etc.) or sound elevations.

In summary, there were several highlights of the proposed localization algorithm.

First, the model's behavior is highly predictable. While previous studies have used k-means clustering in ITD/ILD sound source localization applications (Kim and Okuno, 2013, Ayllon et al., 2012, Davila-Chacon et al., 2013), the data points in their feature space were not confined in the unit circle or form a predictable pattern, which were unique properties of the normalization procedure implemented by Keyrouz's approach.

Second, our model resolved a potential issue in the original Keyrouz methods. As shown in Fig. 4, although there are three clusters at every model frequency presumably corresponding to the three sound sources, it is not obvious which one is which. Keyrouz and colleagues said the same source can be maintained across different frequencies because the center position did not change much between adjacent frequencies. Even when that assumption holds, it will still take some "manual efforts" to pair the clusters with their origins, because they never examined the locational change in a systematical way and did not present any spiral feature.

The model we presented at various frequencies (Fig. 5) can easily resolve this issue. As demonstrated by the two-source example in Fig. 8,

simply comparing the cluster centers with the model constructed at the same frequency can quickly and accurately identify the source location of each data cluster. Our model is even more intriguing in understanding the challenging situation when sound localization is performed under room reverberations.

Third, a major issue in applying some of the previous strategies to real-life hearing devices is the computational complexity that involves heavy information processing and machine-learning algorithms. In Keyrouz's original algorithm, phase normalizations and cluster identifications will have to be performed separately over a large range of frequencies to reconstruct the HRTFs, which are functions of sound frequencies. Next, their matching algorithm involves self-splitting competitive learning and Bayesian fusion to derive a source location in both azimuth and elevation, which is not realistic for real-time hearing devices. Their algorithm would also require the storage of an entire HRTF database.

In the model we proposed here, a single model frequency (such as the 655Hz) can be used to predict the location of a sound source, as long as energy is present at that frequency. This model is also easier to be stored in a hearing device than a complete HRTF database. Our simple classification algorithm can determine the locations of multiple sound sources by mapping the extracted k-means cluster locations to the spiral model normalized to the unit circle. The model performance was largely invariant with the type of sound profiles, except at high frequencies where speech's energy is scarce. This is an important feature because, in real life, sound types can vary all the time. We would not want an algorithm to change according to the ongoing sound type.

However, one should be careful when choosing the frequency, because although low frequencies generally create fewer errors than high frequencies (Fig. 9), we did not propose "the lower, the better". When the frequency is too low, the clusters are highly packed in a small area (Fig. 5A). In fact, under difficult listening conditions, such as room reverberations, the best performance was achieved with the 600 Hz model, because it extends the spiral to almost one full turn. In addition, if there is not sufficient energy present at a particular frequency, a different frequency should be selected.

One limitation in the present study is that, in order to simplify the algorithm, it is required to know k a priori, therefore it would not be ideal for a real-world application where this information is not readily available. This challenge can be overcome by utilizing the method discussed by Keyrouz and colleagues where k is automatically determined using the "self-splitting competitive learning (SSCL)" concept. This approach uses the powerful cluster classifying algorithm based on One-Prototype-Take-One Cluster (OPTOC), meaning one cluster in the feature space is represented by exactly one prototype only (Keyrouz, 2017). Future studies can be improved by attempting to replicate this approach. There are instances however, where this is no longer an issue, when users only want to know a couple, rather than all sources. Finally, due to the lack of energy for human speech at high frequencies, our model is not able to form precise clusters, resulting in limited results at high frequencies for real life applications. We aim to expand this study through finding a solution to this challenge.

Regarding the integration-window size used for STFTs, we used 80 ms to create the broadband model so that the data scattering was largely reduced (Fig. 5A). This window directly relates to the real-time delay of the hearing device. Of course, there is always a tradeoff between the data accuracy and time delay. We suggest using 80 ms in the model creation since it will be performed offline. During the real-time application, shorter window sizes may be considered to reduce the time delay.

4.2. Anechoic condition

In experiment I, to study how a cluster's position changes according to its location in azimuth, an anechoic listening environment with one or more active sound sources was simulated; the effect of room reverberations was ignored. Here, the virtual distance of each sound source was fixed as 1 m, according to the distance of the CIPIC database. Using a single broadband noise, the clusters' center positions followed a spiral pattern. Because this pattern did not vary significantly with sound type or elevations, we chose the single-source broadband pattern as our "model". Due to changes in ITD and ILD dominance that affects spirality, the spiral feature varied with frequency.

Note that the identification of the cluster center for a single source should not be achieved by computing the geometrical *mean* of all the data points on the feature space. We are looking for where the data points are "densest" by running the k-means approach with parameter "1" that specifies the number of clusters, which will not be affected by outliers.

After creating the broadband model, we applied the model to two sound sources with different spatial locations mixed together and tried to predict the azimuth of the fixed and moving sound sources with only one frequency. This is done by comparing the positions of the clusters' centers to the model at the same frequency. Again, the cluster locations were determined by running the k-means approach with parameter "2" that specifies the number of clusters. When both source locations were at 0° elevation, the identifications of both sources matched their predictions using the single-source model (Fig. 7A).

For most of the simulations, we intended to predict sound-source locations in the frontal hemifield using one elevation value (0°). As can been seen in Fig. 7 (B and C), changing the elevations to 45° below (the lowest degree measured in the CIPIC database) and 62° above would slightly deviate from the 0° model (black circles). However, we consider those deviations acceptable and prefer not using multiple elevation models since it was meant to be a simple and fast localization model.

Note that a widely used localization model was the GCC-PHAT model (Knapp and Carter, 1976, Ollivier et al., 2019). We showed that although it works generally fine with broadband noise, it would not perform as well as our spiral model when a short (e.g., 0.15 s) speech sound was the sound source. This is presumably due to the fact that the estimation of a single time delay may not be robust enough for a complex daily listening condition when the energy is sparse.

4.3. Reverberation condition

In experiment II, reverberations were added to a single speech signal. Room impulse responses were obtained from the Aachen Impulse Response (AIR) database (Jeub, 2019, Jeub and Vary, 2009). AIR contains binaural room impulse responses (BRIR) for different listening environments, and two environments were used here: *Stairway*, and *Aula Carolina*. We compared the filtered sounds to the model that was previously found using the CIPIC database.

For the *Stairway*, it can be seen that even though the room type had moderate reverberation ($RT_{60} = 0.83\ s$), the clusters were close to the expected clusters from the model, and the best performance was achieved with the model frequency, 600 Hz. Next, we looked at a larger room with stronger reverberations $(RT_{60} = 5.16\ s.)$, *Aula Carolina*. This listening environment has strong reverberation due to a high ceiling and architecture of the church. Additionally, the speaker was very far from the microphones (9.8 ft). Therefore, as it was expected, the clusters did not fall on the right place in the model. Because of high reverberation and the distance between listener and speaker, the vector strength was significantly reduced, and the resulting cluster was always close to the center.

In fact, a common observation under reverberant conditions is that the vector strength of the cluster center is always reduced (Figs. 11 and 12) compared with the anechoic responses. The more reverberations, the closer the response to (0,0) in the feature space. This phenomenon reflects the smearing effect of one or more echoes on the original sound source, making neither ITD nor ILD cue as prominent as before. How-

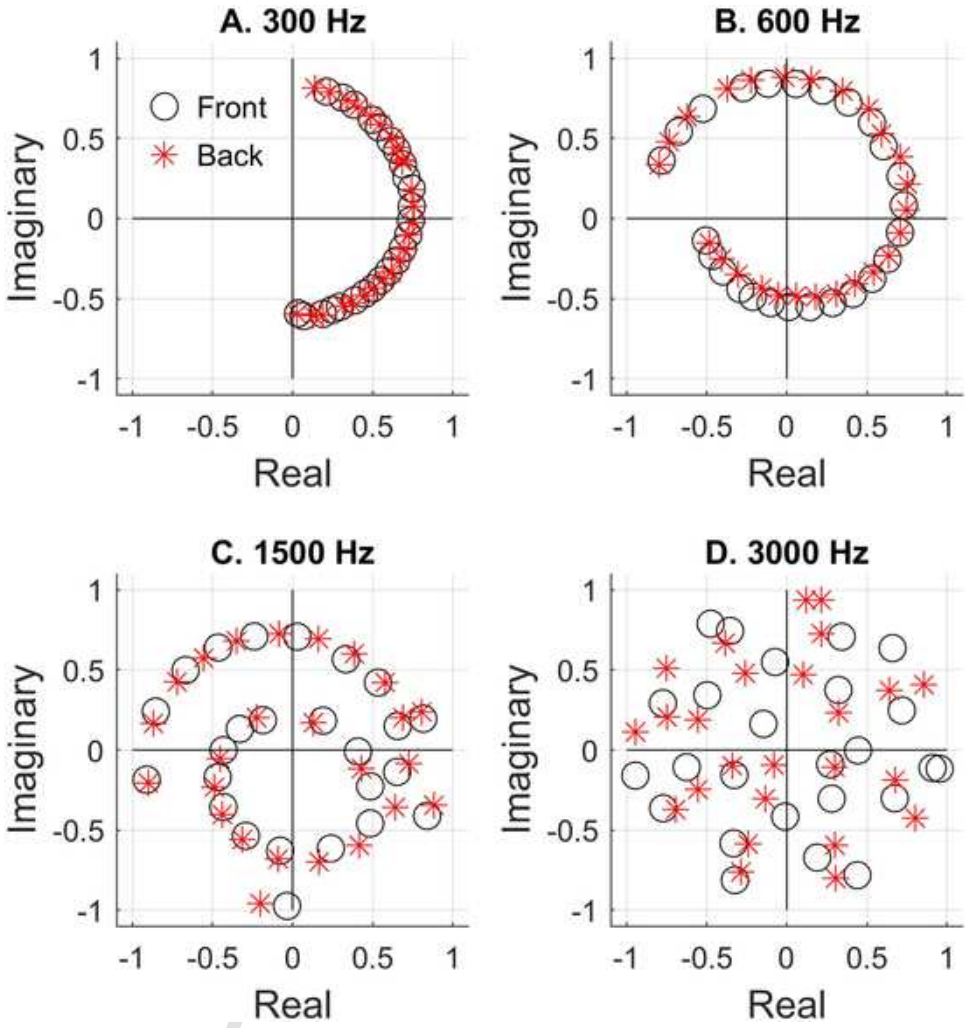

Fig. 14. Repeating the broadband model (Fig. 5) with both front and back sound locations. The model's behavior clearly shows that the localization algorithm based on only one frequency, instead of the entire HRTF, cannot resolve front-back confusions.

ever, the fact that low-frequencies generally performed better than high-frequency models indicates that ITD remains more resilient than ILD under reverberations.

In the original study by Keyrouz, 2017), they added simulated echoes to anechoic sound, and the reverberations interacted with the original sound signal with a signal-to-noise ratio of 20 dB. They achieved an averaged angular error around 10° in both azimuth and elevation for multiple concurrent sources. In the AIR database (Jeub et al., 2009, Jeub, 2019) we used, true reverberations in various types of rooms; RT_{60} was used to quantify the reverberation's strength, which is the time it takes for the sound to attenuate by 60 dB after its termination. Therefore, reverberations we created using the AIR database did not have a constant signal-to-noise ratio. Because our sound stimuli lasted for a long period of time (more than 10 s), the *Aula Carolina* condition with $RT_{60} = 5.16\ s$ would definitely keep evoking overlapping echoes of reflected signals with an effective signal-to-noise ratio much lower than 20 dB.

4.4. Localization of moving sound

Roman and Wang (2008) developed a multi-stage algorithm to track moving sources using binaural input. This algorithm employed a binaural cross-correlation approach to identify spectral regions that reliably separated sources, followed by a Hidden Markov Model (HMM) to identify likely transitions between different configurations of source location and interference. Similarly, a system developed by Dietz et al. (2011) uses ITD and ILD cues as well as a measure of instantaneous interaural envelope phase coherence to create a time-frequency mask that could localize up to five sources in the front field with a maximum error of less than 5°. The effect of head rotation in sound localization of static sound sources is also investigated by Hambrook et al. (2017), which is equivelent to moving sound sources. This study relies on mobile microphone arrays and is not necessarliy applicable to binaural hearing aids.

Our algorithm competes well here, with the smallest average error being 4.5°, and RMS error being 6.4°. Notably their approach also incorporated particle filtering to track sound source motion. Our model addresses moving sounds by constantly updating the listening envrionemnt, such that moving sounds are accounted for at near real-time accuracy with low computational requirements. Since we are only computing a short timeframe but continuously updating it, any moving sound will experience a small lag in localization time.

4.5. Front-back confusions

As mentioned earlier, front-back confusions are inherent issues when sound localization only depends on ITD and/or ILD cues. The only possibility of resolving the issue is to take into account the entire HRTFs. Therefore, the simplified model presented here inevitably suffers from front-back confusions. In Fig. 14, we repeated the broadband

model and superimposed sound locations from the back hemisphere with the frontal hemisphere. It is impossible to distinguish back locations for any of the model frequencies. Therefore, this model only works for the frontal field (or the back, in a similar way).

Overall, our present study has the potential of providing sound source localization for hearing aids, in that it enables the development of a much simpler and faster algorithm to determine the horizontal location of multiple concurrent sound sources. Our results are also on par with the previous HRTF-based studies that used similar methods but required more computational power to support a more complex algorithm. Our model takes a simple approach, while providing effective sound localization performance. The simplicity of our model allows it to be stored in a hearing device, as we do not require a complete HRTF database. Additionally, the developed spiral model is almost invariant to various types of sound (male, female, music, etc.) and sound elevations.

CRediT authorship contribution statement

Jakeh Orr : Visualization, Writing – original draft, Visualization. **William Ebel :** Conceptualization, Methodology, Supervision. **Yan Gai :** Investigation, Software, Writing – review & editing.

Data availability

Data will be made available on request.

References

Zhong, X., Yost, W.A., 2017. How many images are in an auditory scene? J. Acoust. Soc. Am. 141 (4), 2882.

Keller, C.H., Takahashi, T.T., 2005. Localization and Identification of concurrent sounds in the Owl's auditory space map. J. Neurosci. (45), 10446–10461. . p. 25.

Saberi, K., et al., 1991. Free-field release from masking. J. Acoust. Soc. Am. 90 (3), 1355–1370.

Loiselle, L.H., et al., 2016. Using ILD or ITD cues for sound source localization and speech understanding in a complex listening environment by listeners with bilateral and with hearing-preservation cochlear implants. J. Speech Lang. Hear. Res. 59 (4), 810–818.

Saunders, aJMK, G.H., 1997. Speech intelligibility enhancement using hearing-aid array processing. J. Acoust. Soc. Am. 102 (3), 1827–1837.

Makino, S., Lee, T.W., Sawada, H., 2007. Blind Speech Separation. Springer, The Netherlands.

Mills, A.W., 1972. Foundations of Modern Auditory Theory. Vol. II. Academic, New York.

Tollin, aYTC, D.J., 2009. Sound localization: neural mechanisms. Encycl. Neurosci. 137–144.

Musicant, A.D., Chan, J.C., Hind, J.E., 1990. Direction-dependent spectral properties of cat external ear: new data and cross-species comparisons. J. Acoust. Soc. Am. 87 (2), 757–781.

Tollin, D.J., Koka, K., 2009. Postnatal development of sound pressure transformations by the head and pinnae of the cat: monaural characteristics. J. Acoust. Soc. Am. 125 (2), 980–994.

Gardner, M.B., Gardner, R.S., 1973. Problem of localization in the median plane: effect of pinnae cavity occlusion. J. Acoust. Soc. Am. 53 (2), 400–408.

Halupka, NJM, D., Aarabi, P., Sheikholeslami, A., 2005. Robust sound localization in 0.18 /spl mu/m CMOS. IEEE Trans. Signal Process. 53 (6), 2243–2250.

Macdonald, J.A., 2008. A localization algorithm based on head-related transfer functions. J. Acoust. Soc. Am. 123 (6), 4290–4296.

Wang, L., et al., 2016. An iterative approach to source counting and localization using two distant microphones. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (6), 1079–1093.

Rothbucher, M., et al., 2012. HRTF-based localization and separation of multiple sound sources. In: Proceedings of the IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication.

Keyrouz, F., 2017. Robotic binaural localization and separation of multiple simultaneous sound sources. In: Proceedings of the IEEE 11th International Conference on Semantic Computing (ICSC). pp. 188–195.

Keyrouz, F., 2014. Advanced binaural sound localization in 3-D for humanoid robots. IEEE Trans. Instrum. Meas. 63 (9), 2098–2107.

Keyrouz, F., 2015. Binaural range estimation using head related transfer functions. In: Proceedings of the IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI). pp. 89–94.

Keyrouz, F., 2008. Efficient Binaural Sound Localization for Humanoid Robots and Telepresence Applications (Ph.D. thesis). Technische Universitäat München.

Calandruccio, L., Smiljanic, R., 2012. New sentence recognition materials developed using a basic non-native English lexicon. J. Speech Lang. Hear. Res. 55 (5), 1342–1355.

Gelfer, M.P., Mikos, V.A., 2005. The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. J. Voice 19 (4), 544–554.

Algazi, V.R., et al., 2001. The CIPIC HRTF database. In: Proceedings of the IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575). pp. 99–102.

Jeub, M., Schafer, M., Vary, P., 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In: Proceedings of the 16th International Conference on Digital Signal Processing. pp. 1–5. p.

Jeub, M., AIR Database. https://www.mathworks.com/matlabcentral/fileexchange/29073-air-database, MATLAB Central File Exchange. 2019.

Knapp, C., Carter, G., 1976. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust. Speech Signal Process. 24 (4), 320–327.

Ollivier, B., et al., 2019. Noise robust bird call localisation using the generalised cross-correlation with phase transform in the wavelet domain. J. Acoust. Soc. Am. 146 (6), 4650.

Liu, KR, J., 2007. Multiple self-splitting and merging competitive learning algorithm, Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, p. 8.

Lyon, R., 1984. Computational Models of Neural Auditory Processing. International Conference on Acoustics, Speech, and Signal Processing.

Baumann, C., Rogers, C., Massen, F., 2015. Dynamic binaural sound localization based on variations of interaural time delays and system rotations. J. Acoust. Soc. Am. (635), 138.

Raspaud, M., Viste, H., Evangelista, G., 2009. Binaural source localization by joint estimation of ILD and ITD. IEEE Trans. Audio Speech Lang. Process. 18 (1), 68–77.

Roman, N., Wang, D., Brown, G.J., 2003. Speech segregation based on sound localization. J. Acoust. Soc. Am. 114 (4 Pt 1), 2236–2252.

Kraljević, L., et al., 2020. Free-Field TDOA-AOA Sound Source Localization Using Three Soundfield Microphones, 8. IEEE Access, pp. 87749–87761.

Dang, X., Cheng, Q., Zhu, H., 2019. Indoor multiple sound source localization via multi-dimensional assignment data association. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (12), 1944–1956.

Grumiaux, P.A., et al., 2022. A survey of sound source localization with deep learning methods. J. Acoust. Soc. Am. 152 (1), 107.

Vecchiotti, P., Ma, N., Squartini, S.e.a., 2019. End-to-end binaural sound localisation from the raw waveform. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, UK.

Wang, Z.Q., Zhang, X., Wang, D., 2019. Robust speaker localization guided by deep learning-based time-frequency masking. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (1).

Ma, N., May, T., Brown, G.J., 2017. Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (12), 2444–2453.

Kim, U.H., Okuno, H.G., 2013. Robust localization and tracking of multiple speakers in real environments for binaural robot audition. In: Proceedings of the 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS). IEEE, Paris, France.

Ayllon, D., et al., 2012. Speech Source Separation Using a Generalized Mean Shift Algorithm. Elsevier Signal Processing.

Davila-Chacon, J., et al., 2013. Neural and statistical processing of spatial cues for sound source localisation. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN). IEEE.

Jeub, MS, M., Vary, P., 2009. A binaural room impulse response database for the evaluation of dereverberation algorithms. In: Proceedings of the 16th International Conference on Digital Signal Processing. pp. 1–5.

Roman, N., Wang, D., 2008. Binaural tracking of multiple moving sources. IEEE Trans. Audio Speech Lang. Process. 16.

Dietz, M., Ewert, S., H, V., 2011. Auditory model based direction estimation of concurrent speakers from binaural signals. Speech Commun. 53 (5), 592–605.

Hambrook, D.A., et al., 2017. A Bayesian computational basis for auditory selective attention using head rotation and the interaural time-difference cue. PLoS One.