IWSSIP 2010 - 17th International Conference on Systems, Signals and Image Processing

Spatial Audio to Assist Speaker Identification in Telephony

Konrad Blum, Gert-Jan van Rooyen and Herman A. Engelbrecht
Department of Electrical and Electronic Engineering
Stellenbosch University, South Africa
Email: kblum@ml.sun.ac.za, gvrooyen@sun.ac.za, hebrecht@sun.ac.za

Abstract—Telephony has developed substantially over the years, but the fundamental auditory model of presenting audio monaurally has not changed since the telephone was first invented. Monaural audio is very difficult to follow in a multiple-source situation such as a conference call. We believe that it could be beneficial to a participant in conference call to know who is currently speaking. This paper evaluates the benefit of a spatial audio telephony application by comparing three spatial audio models against monaural audio. Experiments in which subjects had to identify the active speaker out of a number of possibilities, demonstrate that spatial audio affords the user an approximately two-fold increase in correct speaker identification rate in a conference call. These results demonstrate that spatial audio is easier to follow than monaural audio and show that spatial audio in telephony can aid speaker identification in conversations with multiple speakers.

Keywords-spatial audio;telephony;psychoacoustics.

I. Introduction

Sound originating from a specific point in space will travel a slightly different path to each ear and the human brain processes these spatial cues to locate sounds in space [1]. This spatial information allows a listener to focus their attention on a single speaker in an environment where many different sources may be active at the same time; this is known as the "cocktail party effect" [2]. It is possible to reproduce these spatial cues in a sound recording using techniques such as head-related transfer functions [3] and binaural recording [4] to allow a listener to hear localized audio, even when sound is reproduced through a headset.

Although the use of spatial audio is in relatively widespread use in e.g. the entertainment sector, it is not a common feature in telephony. Modern telephony generally makes use of the basic monaural audio model of mixing the audio streams of different participants in a conference call together, discarding the spatial information [5]. Making use of a person's ability to separate sound based on perceived spatial location allows one to better communicate information than is possible with monaural communication [6]. Knowing who is currently speaking in a conference call can be important for a user, as one would answer a question asked by the lead developer during a job interview differently than one asked by a member of human resources. In this paper, we describe the use of a VoIP system that utilizes various spatial audio models [7] and provide experimental evidence of its effectiveness in assisting a user in speaker identification.

II. Sound Localization

The ability of a listener to determine the range and direction of a sound is known as sound localization. If a sound source is not directly to the front of a listener then the sound will follow a different path to each ear, with one path experiencing a longer delay. This results in an interaural time difference (ITD), which can range between 0 μs (for a source on the median plane) and 650 μs (for a source directly to one side) [1]. This longer propagation distance and shadowing by the head and torso will also attenuate the sound traveling to the further ear more, causing an interaural level difference (ILD). The brain uses these binaural cues to locate sounds in space. Diffraction around the head reduces shadowing for low frequencies, making the ILD frequency dependent which provides the brain with more localization cues. The spectral content of the sound reaching the eardrum is further changed by resonance and reflections due to the shape of the pinna (the visible part of the outer ear) depending on the angle of incidence [4]. Localization accuracy is dependent on the spectral content of the sound source [8]. The combination of the ITD, ILD and acoustic filtering effect caused by the pinnae, head and torso gives us our ability to localize sounds [9].

III. Spatial Audio Rendering Techniques

A. Head-Related Transfer Functions

As discussed in Sec. II, the spectral filtering resulting from the different paths that a sound source from an arbitrary point in space travels to reach each ear provides cues that aid a listener in locating the sound source. This spectral filtering can be expressed by head-related transfer functions (HRTFs) as shown in Fig. 1. HRTFs can be measured, and such databases [10], [11] can be used to give sound the perception of direction by convolving a monaural audio source $x(t)$ with the impulse response corresponding to the HRTF pair, giving outputs for the left and right ear respectively

$$x_L(t) = x(t) \star h_L(t) \tag{1}$$

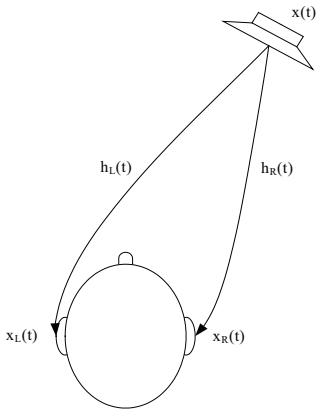

Fig. 1. Head-related transfer functions express the spectral filtering due to the different path traveled by a sound source to each ear.

and

$$x_R(t) = x(t) \star h_R(t), \tag{2}$$

with $h_L(t)$ and $h_R(t)$ being the impulse responses for the left and right ear respectively. The HRTFs of every person are unique and the brain undergoes a constant calibration process to ensure accurate sound localization [12]. HRTFs from the Listen database [10] were used in this project.

B. Stereo Panning

Stereo panning is a technique in which a monaural signal is placed in a stereophonic sound field, and setting the apparent horizontal position of the sound by changing the output levels of the two loudspeakers. The most common of these, the "sine-cosine" pan law [13] relies on loudspeakers that are placed 45° to the left and to the right of where the listener is facing to, which is not the case with headphones. Extending the stereo panning model to work for headphones (or loudspeakers placed 90° to the left and to the right of the listener) and placing the direction that the listener is facing at 0° gives

$$g_L = \cos(\theta/2 + 45^\circ) \tag{3}$$

and

$$g_R = \sin(\theta/2 + 45^\circ), \tag{4}$$

where g_L and g_R are the gain factors for the left and right ear respectively and the panning angle θ is as shown in Fig. 2 [7]. The audio is played only to the subject's left ear when $\theta = -90^\circ$, to both ears equally when $\theta = 0^\circ$ and only to the right ear when $\theta = 90^\circ$. The headphones model maintains constant energy (and therefore constant loudness) because $g_L^2 + g_R^2 = 1$.

C. Basic Binaural Model

The basic binaural model is a pure geometric model that is developed as a simplification of the HRTF model discussed in Sec. III-A. The model uses the different distances traveled by sound to each ear, as can be seen in Fig. 1, to calculate ILD and ITD functions. The model does not take into account any reflection, absorption and diffraction effects resulting from the subject's head and

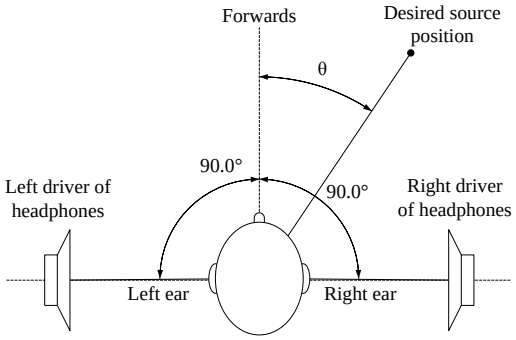

Fig. 2. The positions of the headphone drivers and desired source position relative to the listener.

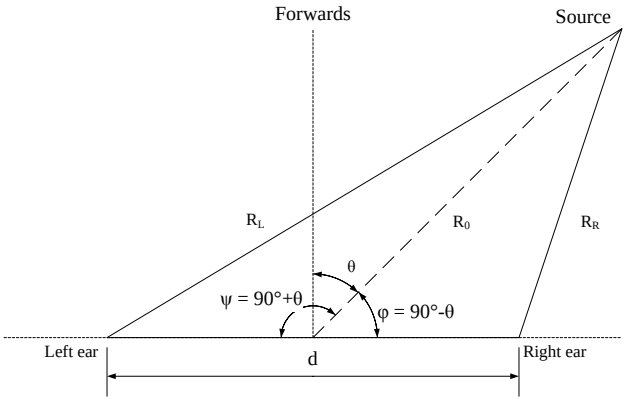

Fig. 3. Geometry of basic binaural model, looking from the top.

torso. From the geometry shown in Fig. 3 [7], the distance from the source to the left ear is

$$R_L = \sqrt{R_0^2 + \frac{d^2}{4} - R_0 d \cos(90^\circ + \theta)} \tag{5}$$

where R_0 is the distance from the source to the center of the listener's head and d the inter-aural distance (spacing between the two ears) [8]. Since sound attenuates over distance according to the inverse square law [9], the gain at a distance R_L from the source is

$$g_L = \left(\frac{R_0}{R_L}\right)^2. \tag{6}$$

The time delay resulting from sound propagating to the left ear is $\tau_L = R_L/c$, where c is the speed of sound. Combining the gain (ILD) and time delay (ITD) functions, the signal presented to the left ear is

$$x_L(t) = \frac{g_L x(t - \tau_L)}{g_{max}}, \tag{7}$$

with the maximum gain, $g_{max} = R_L(\theta = -90^\circ)$. The signal presented to the right ear is calculated in a similar fashion.

IV. Speaker Identification

We hypothesize that spatial audio affords a listener greater ability in identifying an active speaker in a multiple speaker situation than monaural audio.

A. Experimental Protocol

Two experiments were designed to emulate a conference call scenario, with participants that are unfamiliar to the user, but where it is important to know who is speaking. We chose to take an off-line approach and use pre-generated audio files instead of testing using a live system to ensure repeatability with different test subjects [7]. The subjects had to use headphones to listen to audio samples, each with a single sentence being spoken by a single speaker and identify which speaker was active. The first experiment compared only HRTF spatialization to monaural audio and featured four possible speakers. The second experiment compared all three techniques discussed in Sec. III against monaural audio with six possible speakers. The sentences were chosen from the Grid audiovisual speech corpus [14].

1) Four Possible Speakers: Ten sets of audio files were generated, with each set containing four files, each with sentences from only a single speaker. There is no temporal overlap between sentences within a single set. The segments in each set consist of two stages: the introduction stage and the identification stage. The introduction stage has two sentences from each speaker, for a total of eight sentences, in order to allow the subjects to familiarize themselves with the voices and spatial positions of each of the speakers. The identification stage has four sentences from each speaker in a random order. The subjects are only required to identify the speakers in the second stage which requires 160 identifications per subject. Each test consisted of five monaural and five spatial audio files, alternating between monaural and spatial, with half of the subjects starting on a monaural file and the other half on a spatial file. Each test subject had to listen to the test files in the order provided and attempt to identify which speaker was active for each of the speech segments. The speakers were placed at azimuths of $285°$, $330°$, $30°$ and $75°$, where $0°$ is in the direction that the listener is facing.

2) Six Possible Speakers: In this experiment, subjects were presented with audio samples, each with a single utterance from a single speaker, and had to identify which of the six speakers was active. The speakers were placed at $300°$, $330°$, $30°$, $60°$, $120°$ and $240°$. The subjects were again given the opportunity to familiarize themselves with the speakers.

B. Results and Discussion

1) Four Possible Speakers: Sixteen subjects took part in the experiment. If a subject made a mistake, either missing a sentence or identifying too many sentences, all the data from that particular run was discarded. A total of six mistakes were made. This left 76 valid monaural runs totalling 1216 valid monaural identifications and 78 valid spatial runs totalling 1248 valid spatial identifications. Fig. 4 shows the speaker identification rates averaged across all the subjects for each of the monaural and spatial runs, with the probability of guessing correctly included for comparison purposes. On average, the subjects identified the source correctly 43% of the time with monaural audio and 88% of the time with spatial audio.

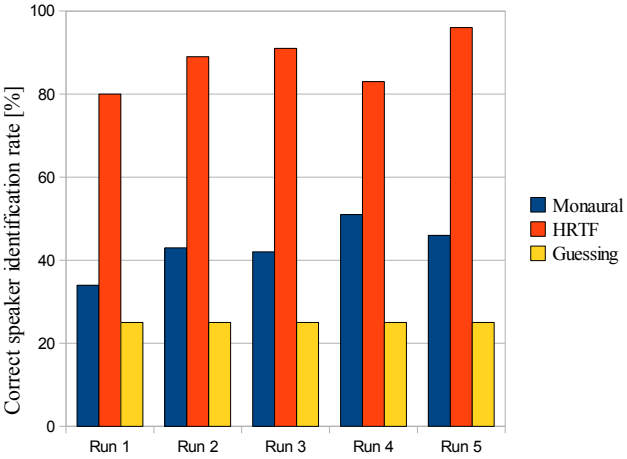

Fig. 4. Speaker identification rates with four possible speakers.

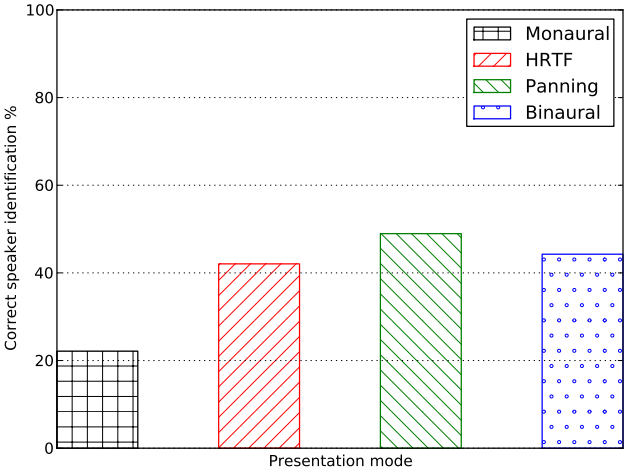

Fig. 5. Speaker identification rates with six possible speakers.

2) Six Possible Speakers: As the experiment was done with potentially anonymous subjects, some cleaning of the collected data was necessary by dropping all data from subjects who did not complete enough test runs. A total of 62 subjects took part in the experiment, with 45 subjects completing enough test runs; 93.27% of the collected data was used.

Fig. 5 shows the average speaker identification rate for each auditory presentation mode. The spatial presentation modes greatly surpass the monaural presentation, with the HRTF spatial and binaural modes achieving scores just less than twice that of the monaural mode. The headphone panning model performs the best out of the three spatial models, with more than twice the identification rate of the monaural mode. The overall scores being lower than those observed in the previous experiment can be largely attributed to the larger number of possible source positions, but an approximately two-fold increase in identification rate of spatial audio over monaural audio is once again apparent.

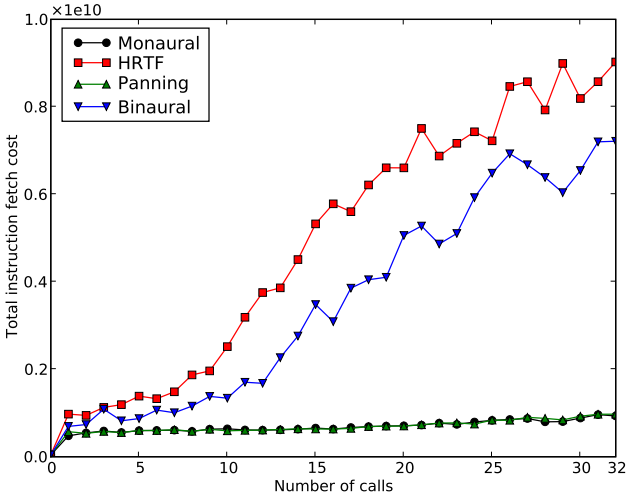

Fig. 6. The total instruction cost for each presentation mode.

V. Resource Usage

The three spatial audio techniques discussed in Sec. III were implemented in a VoIP application [7], a real-time scenario in which resource usage needs to be managed. The change in required processing power resulting from the implementation of spatial audio compared to monaural audio was investigated. The application was profiled using Valgrind [15]. The total instruction cost as a function of the number of active calls was determined for each auditory presentation mode and is shown in Fig. 6. Each mode has a large cost for the first call, which can be attributed to initialization of the networking and audio modules, with additional costs of subsequent calls increasing in a somewhat linear fashion. On average, HRTF spatialization costs 9.03 times as much as the monaural audio, panning 1.03 times and binaural 6.86 times.

VI. Conclusion

The experiments detailed above show that spatial audio provides a listener with an approximately two-fold increase in active speaker identification rate in a conference call situation. Headphone panning has more dramatic energy distribution than the other spatial audio, with a source panned completely to the left presenting all the energy to the left ear, which can explain why headphone panning outperforms the HRTF spatialization and the basic binaural model in speaker identification tasks. The high performance and the fact that the headphone panning model also only uses 3% more processing power than the monaural mode, in contrast to the more expensive HRTF spatialization and binaural models, makes it the best choice for tasks requiring speaker identification.

The experiment was conducted using subjects that were not familiar with the system or the specific HRTFs in use. Correct identification rates are expected to improve as the user becomes more familiar with the HRTFs [12] or when using HRTFs from a subject with anthropometric measurements similar to their own. In conclusion, this research has shown the benefit of replacing the current monaural telephony model with a spatial audio model that attaches perceptual direction to each source.

VII. Future Work

A key component to the performance of the system in terms of presenting the correct perceived spatial location of the audio source to the user is dependant on the quality of the HRTFs that are used. HRTFs from a downloadable database were used for the project and are non-individualized because they were not measured using the subject's own auditory system. Research shows that such non-individualized HRTFs can lead to errors in localization, especially with respect to elevation and front-to-back discrimination [16]. Performance can be improved by using HRTFs that are either measured directly from the subject [17] or by individualizing generalized HRTFs using anthropometric measurements [18].

Acknowledgement

The authors would like to thank MIH Holdings for funding the research.

References

[1] M. A. Akeroyd, "The psychoacoustics of binaural hearing," *International Journal of Audiology*, vol. 45, pp. 25–33, 2006.
[2] B. Arons, "A review of the cocktail party effect," *Journal of the American Voice I/O Society*, vol. 12, pp. 35–50, 1992.
[3] D. S. Brungart and W. M. Rabinowitz, "Auditory localization of nearby sources. head-related transfer functions."
[4] F. Rumsey, *Spatial Audio*. Focal Press, 2001.
[5] Y. Kanada, "Multi-context voice communication in a SIP/SIMPLE-based shared virtual sound room with early reflections," in *NOSSDAV '05: Proceedings of the international workshop on Network and operating systems support for digital audio and video*. New York, NY, USA: ACM, 2005, pp. 45–50.
[6] J. J. Baldis, "Effects of spatial audio on memory, comprehension, and preference during desktop conferences," in *CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems*. New York, NY, USA: ACM, 2001, pp. 166–173.
[7] K. Blum, "Evaluating the applications of spatial audio in telephony," Master's thesis, Stellenbosch University, March 2010.
[8] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization*. MIT Press, 1997.
[9] F. A. Everest, *The Master Handbook of Acoustics*, 4th ed. McGraw-Hill Professional, 2000.
[10] Listen HRTF Database, http://recherche.ircam.fr/equipes/salles/listen/, last accessed: November, 2009.
[11] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.
[12] P. M. Hofman, J. G. V. Riswick, and A. J. V. Opstal, "Relearning sound localization with new ears," *Nature Neuroscience*, vol. 1, no. 5, pp. 417–421, September 1998.
[13] D. Griesinger, "Stereo and surround panning in practice," in *Audio Engineering Society 112th Convention*, 2002, pp. 1–6.
[14] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition."
[15] "Valgrind," http://valgrind.org, last accessed: November, 2009.
[16] E. M. Wenzel, M. Arruda, D. K. Kistler, and F. L. Wightman, "Localization using nonindividualized head-related transfer functions."
[17] V. Algazi, R. Duda, D. Thompson, and C. Avendano, "The CIPIC HRTF database," in *Proc. 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 99–102.
[18] S. Xu, Z. Li, and G. Salvendy, "Improved method to individualize head-related transfer function using anthropometric measurements," *Acoustical Science and Technology*, vol. 29, no. 6, pp. 388–390, 2008.