Literature Review of the Use of Binaural Audio in Speech Processing

David S. Asgrimsson
Department of Computer Science
KTH
dsas@kth.se

Abstract

This literature review presents the basics of binaural audio and it's uses, especially in speech localisation and in 'cocktail party' problems. Binaural audio is developed to closely resemble how the human ear naturally captures sounds, with regards to how the human brain is adapted to hearing in a three dimensional environment. The most common method in speech processing is to use stereo audio, which does not take the biological factor of hearing into account and as such contains less spatial data. By addressing the filtering and spatial localization functions of the human ear and using binaural audio, we can more concretely address localization and cocktail party problems. This method is promising, but does not necessarily guarantee better performance than multi-microphone set ups.

1 Introduction

Speech processing is the study of speech signals and the processing methods of these signals. The signals are recorded in analog manner and has to be converted to a digital representation. This signal is most often stereo audio, a combination of two separate audio channels to create the impression of sound heard from the left or right directions, or even mono audio, only from one microphone. This method does not directly capture a 3-D soundscape as the human ear naturally does, but a 2-D mapping to a left-right configuration.

Binaural recording is a method of recording sound that uses two microphones, arranged with the intent to create a 3-D stereo sound by addressing how the human ear captures sound. This format is fitted specifically for the human ear, and presents a different sound by mimicking the physical attributes of human hearing, that captures sound much more naturally. Such recording are made with devices such as in figure 2.

Aspects of speech processing includes transfer and output of sound signals. Humans use binaural hearing to localize sound, by comparing the information received from each ear in a signal processing step that involves a synthesis. One of the biggest challenges of a real life speech recognition system is the ability to adapt to a real life noisy environment. Binaural recording enhances the intelligibility of speech in noisy settings, and in the presence of many speakers talking at the same time, often called the *cocktail party problem* (Cherry 1953; Haykin and Chen 2005).

Binaural fusion is a cognitive process that involves the processing of different auditory signals presented to the two separate human ears. In humans, this process is essential in processing speech as each ear picks up sound input that is slightly different and this difference contains essential information. The process of binaural fusion is important for computing the location of sound sources by comparing and fusing the slightly different inputs from the two ears. Sound segregation on the other hand identifies acoustic components that may come from several different sources.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

With a clear distinction between left and right perspectives and biologically inspired methods, binaural sound brings a more authentic sound to the extent that a person listening to such a recording can be convinced that the recorded sound is real and happening first hand, not recorded (Gardner and Martin 1995).

One pitfall of binaural recording technology is that the user needs to be wearing headphones, to get the full effect. Of course the recording can be played in speakers, but the localization effect will be lost. I recommend listening to The Virtual Barbershop to get the full effect.

1.1 Functions of the human ear

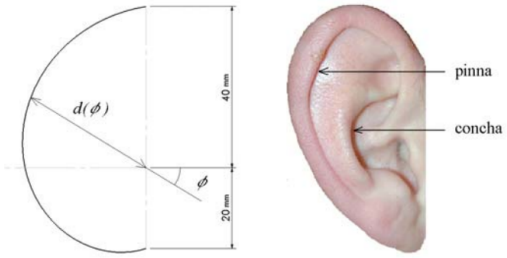

Figure 1: The human outer ear, or auricle and a mathematical model used in robot localization (Hornstein et al. 2006).

In mammals, sound travels through three main parts of the ear to be heard. These are the outer, middle and inner ear. The fundamental biological facts presented in this section are taken from Windelspecht and Sylvia S. Mader 2016.

Outer ear is the part of the ear that we can see called the pinna as well as the inside of the ear called the ear canal. Sound first travels through the pinna and ear canal then to the eardrum at the end of the canal which transfers the air vibration into physical vibrations.

The vibrations continue through to the middle ear and reaches the hammer, anvil and stirrup. They transfer the sound vibrations to the inner ear.

In the inner ear the sound reaches a small shell like tube called the cochlea that contains a fluid, which moves tiny hairs that send electrical signals via nerves directly to the brain.

The auricle or the pinna is the visible part of the ear that resides outside the head. The auricle's functions are to collect sound and transform it. The auricle collects sound and amplifies the sound and directs it to the auditory canal. The filtering effect of the outer ear preferentially selects sounds in the frequency range of human speech Langendijk and Bronkhorst 2002.

Amplification of sound by the pinna, membrane and middle ear causes an increase in level of about 10 to 15 dB in a frequency range of 1.5 kHz to 7 kHz (Hebrank and Wright 1974).

The pinna slightly changes the sound it captures by eliminating a small segment of the frequency spectrum, this band is called the pinna notch. The pinna works differently for low and high frequency sounds. For low frequencies it diirects sounds toward the ear canal but for high frequencies the function is not as well understood and is more complex. While some of the sounds that enter the ear travel directly to the canal, others reflect off the pinna first. These enter the ear canal after a very slight delay. In the affected frequency band, the pinna notch, the pinna creates a band-stop or notch filtering effect. This filter typically indicate that up–down cues are located mainly in the 6–12-kHz band, and front–back cues in the 8–16-kHz band (Langendijk and Bronkhorst 2002). It also is directionally dependent, affecting sounds coming from above more than those coming from straight ahead. This aids in vertical sound localization.

1.2 Stereo audio

Stereophonic sound, more commonly called stereo, is the reproduction of sound using two or more audio loudspeakers. This creates a sound heard from two main directions. In popular usage, stereo usually means two-channel sound recording and sound reproduction using data from two speakers.

This construction is easy in production, but is not a biologically inspired method of capturing sound. This therefore creates a sound that is often missing a 3 dimensional element. The focus is mostly on a 2 dimensional, left right plane.

1.3 Binaural audio

Binaural recording takes the stereo method further by using biological inspired features by placing two microphones in ear-like cavities on either side of a stand or dummy head. See figure 2 for such a product. The dummy head recreates the density and shape of a human head and the latex ears recreate the functions of the human ear. The filtering effect by the pinna is therefore captured. These microphones capture and process sound closely to as it would be heard by human ears, preserving interaural cues. These cues are detailed in chapter 2.

Figure 2: A Neumann KU 100 binaural recording device. A dense human like head contains two omnidirectional condenser microphones placed at a specific distance from each other to imitate the human ears. Realistic latex ears are placed outside the microphones, to imitate the filtering and localization effect of the human ears. (Taken from Neumann product catalog)

2 Use in speech processing

One of the key observations derived from (Cherry 1953) was that it is easier to separate the sources heard binaurally than when they are heard monaurally. Nature gives us two ears to enable us to perceive the dynamic outer world and provide the main sensory information sources. Binaural processing is crucial in certain perceptual activities by comparing the two recieved sounds, such as depth perception and sound localization. Given one sound source, the two ears receive slightly different sound patterns due to delay produced by the physical placement of the ears and head. The brain is known to be very tuned to certain cues of sounds using varieties of acoustic differences perform specific tasks. The slight differences in these cues are sufficient to identify the location and direction of the incoming sound waves.

2.1 Aural cues and head related transfer function

The *interaural time difference* (ITD) is the difference in arrival time of a sound between two ears. It is important in the localization of sounds, as it provides a signal of the direction or angle of the sound source from the head. If a signal arrives at the head from one side, the signal has further to travel to reach the far ear than the near ear. This difference results in a time difference between the sound's arrivals at the ears, which is detected and aids the process of identifying the direction of sound source. See figure 4 for an illustration of the ITD.

An *interaural intensity difference* (IID) is produced because the head blocks some of the energy that would have reached the far ear, especially at higher frequencies. This makes the sound arriving to each ear slightly different, as made clear in figure 3. The brain decodes the IID difference to assess location of the sound source. This is easy to imagine, that the physical head blocks sound coming from a single direction making the sound different to each ear.

The *interaural phase difference* (IPD) refers to the difference in the phase of a wave that reaches each ear, and is dependent on the frequency of the sound wave and the interaural time differences. IPDs are useful as the human ear has the ability to detect differences in phase, and with the combination of IPD and ITD, helps determining where the sound originated from, and also to identify the frequency of the sound.

Wenzel et al. 1993 and others used probe microphones in the ear to measure and describe the transfer function from sound source to eardrum. This transfer function is commonly referred to as the *head-related transfer function* (HRTF), and its time-domain analog is the head-related impulse response (HRIR). HRTF measurements and their inverse Fourier transforms have been a crucial component in a number of systems that attempt to stimulate natural 3-D acoustical environments. This transfer functions imitates the function of the physics of the human ear and head has on sound.

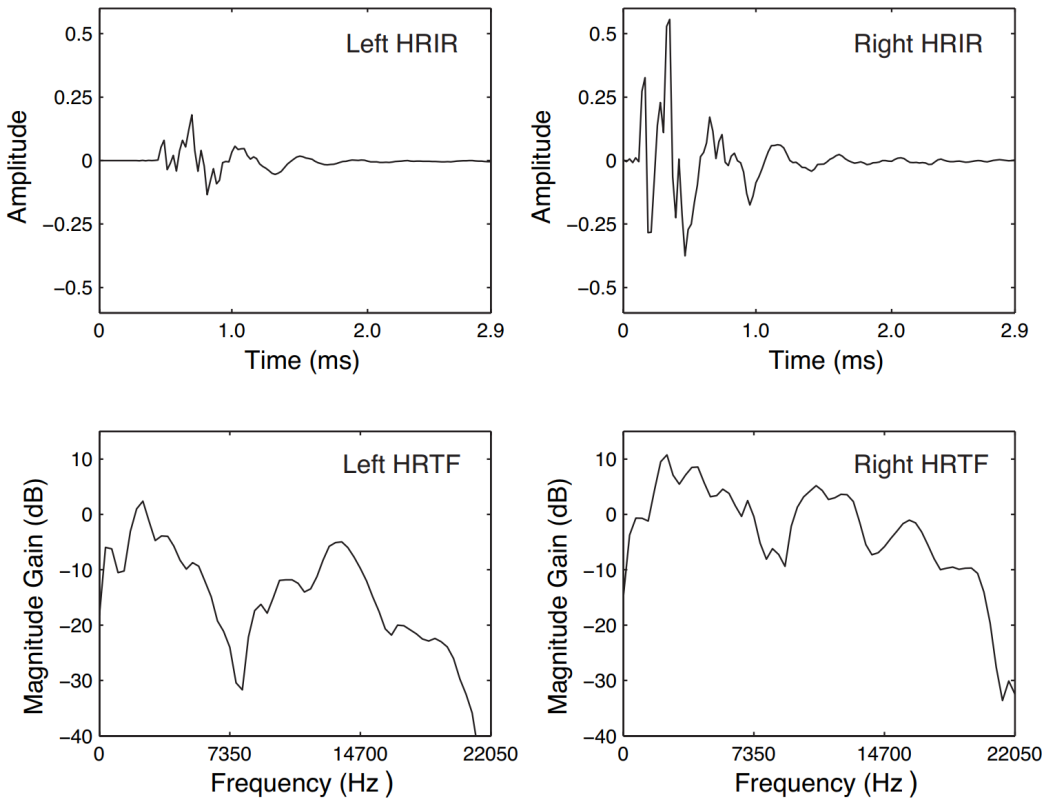

Figure 3: Here the difference between what the left and right ear hears is clearly seen in the HRTFT and HRIR (Gardner and Martin 1995)

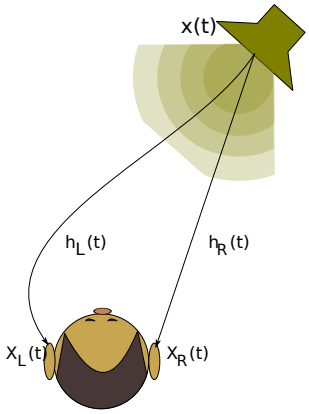

Figure 4: The Head Related Transfer Function is used to imitate the effect that the physical head has on the sound arriving at each ear. This difference is used by the brain to assess localization of sound. Taken from Wikipedia.

2.2 Localization

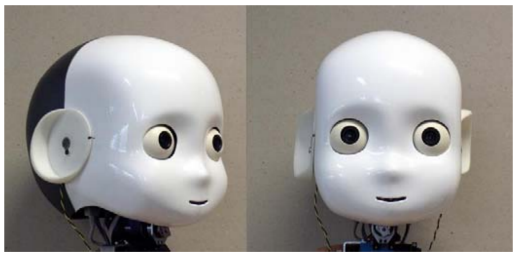

Figure 5: The iCub with spiral shaped ears used for better robot localization (Hornstein et al. 2006).

Extensive studies by a number of researchers have revealed perceptual clues that are useful for sound source localization. They include the interaural level difference, the interaural time difference, and the transformation caused by the outer ear. These clues are implicitly included in the head related transfer function.

Hornstein et al. 2006 present a method by using simple spiral-shaped ears that has similar properties to the human ears to extract spatial information that make it possible to accurately estimate the location of a sound source in both the horizontal and vertical plane using only two microphones and human-like ears. See figure 5. In robotics, notches have been often been left out since they are considered complex and difficult to use. Ears are the main device for humans' ability to estimate the elevation of the sound source, but this has often been compensated by having multiple microphones. The robot is shown to be able to learn its HRTF and build automatically can update its map using vision and compensate for changes in the HRTF due to changes to the ears or the environment

Keyrouz and Diepold 2006 propose a binaural sound source localization technique based on using only two microphones placed inside the ear canal of a robot dummy head. The head is equipped with artificial ears and is mounted on a torso. In contrast methods using microphone arrays, the method employs two microphone and is based on a simple correlation approach using a generic set of head related transfer functions (HRTFs). The proposed method is demonstrated through simulation and is further tested in a household environment. This set up proves to be very noise-tolerant and is able to localize sound sources in free space with high precision and low computational complexity

Deleforge, Forbes, and Horaud 2015 present an algorithm for source separation and localization by finding latent low-dimensional manifold of binaural input. This is done with a Bayesian estimation of the locations and time-frequency masks of several sources. This method could be used to accurately

localize in both azimuth and elevation and separate mixtures of natural sound sources. Comparisons of the proposed approach with several existing methods reveal that the combination of acoustic-space learning with Bayesian inference enables our method to outperform state-of-the-art methods.

Kim, Nakadai, and Okuno 2015 conducted experiments using two dummy heads equipped with small or large pinnae and showed that localization errors were reduced by a large margin on average with the new time delay factor compared with the conventional GCC-PHAT method and better on average over the entire azimuth than with a conventional head related transfer function (HRTF)-based method.

Argentieri, Danès, and Souères 2015 provide a review of state-of-the-art of sound source localization in robotics, specially with a robust performance. Binaural techniques aim at reproducing artificially the human auditory system, but the difficulty to exploit elementary acoustic cues has been underlined, together with the fundamental role of the head in the localization process. Though several models have been proposed in the literature, the most basic of them are not sufficient to explain experimental measurements in an anechoic room. On the other hand, array processing techniques involving a larger number of microphones turn out to be more accurate and robust.

2.3 Talker identification and Cocktail party problems

Drullman and Bronkhorst 2000 describe a study on the possible merits of binaural sound through headphones for bandlimited speech with respect to intelligibility and talker recognition against a background of competing voices. Average results for 12 listeners show an increase of speech intelligibility for binaural presentation for two or more competing talkers compared to conventional presentation.

Hawley, Litovsky, and Culling 2004 study the *cocktail party problem* using virtual stimuli. Speech reception thresholds were measured for sentences presented from the front in the presence of one, two, or three interfering sources. For a single interferer, there was a binaural advantage of 2–4 dB for all interferer types. For two or three interferers, the advantage was 2–4 dB for noise and speech-modulated noise, and 6–7 dB for speech and time-reversed speech. These data suggest that the benefit of binaural hearing for speech intelligibility is especially pronounced when there are multiple voiced interferers at different locations from the target, regardless of spatial configuration; measurements with fewer or with other types of interferers can underestimate this benefit.

Roman and Wang 2008 presented a binaural method for tracking multiple moving sources. Binaural cues are assumed to be strongly correlated with source locations in time-frequency regions dominated by only one source which describes the azimuths of all active sources at a particular time frame. This assumption is used to extend a hidden Markov model for multipitch tracking to the domain of multi-source localization and tracking.

Bronkhorst 2015 summarizes widespread research in psychoacoustics, auditory scene analysis, and attention, all dealing with early processing and selection of speech often coined as the *cocktail party problem*. They find that sounds can be grouped and selected using primitive features such as spatial location and fundamental frequency. Binaural perception and sound localization partly rely on the same cues, which indicates that there is an overlap in the processing at peripheral and brainstem levels. This means that it is still not clear to what degree binaural speech perception depends on one's ability to localize sound target and interfering sound sources.

3 Implementation and evaluation

The general results from this literature review is that when designing a system that deals with many interacting speakers, or the localization of sound, it is favorable to consider the functions of the human ear. Using binaural technology will most likely be beneficial over using a mono, or stereo system. A multi-microphone solution is generally considered to be better than binaural.

To implement this, consider using two omnidirectional microphones located at a similar distance to the human ears. The two microphones should then be fitted with an ear-like device, to aid in localization of the sound as shoen in Hornstein et al. 2006.

This is a relatively simple procedure, and can have a considerable effect. To evaluate the system, you could compare the results from a mono/stereo solution to the binaural system as is done in Hawley, Litovsky, and Culling 2004

4 Conclusion

This literature review goes over the fundamental values of using binaural audio for talker identification and localization. This area of research is well defined and has been active for many decades. For many uses, the imitation of the physical features of human hearing result in better performance, especially in the spatial localization of sound. This is achieved by using binaural recording methods. Using binaural methods in stead of multiple microphone sources can be more complex, and not necessarily better. This makes the problems stated in the review open for more research, as the binaural method is very often beneficial, but not always applicable.

References

Argentieri, Sylvain, Patrick Danès, and Philippe Souères (2015). "A survey on sound source localization in robotics: From binaural to array processing methods". In: *Computer Speech & Language* 34.1, pp. 87–112.

Bronkhorst, Adelbert W (2015). "The cocktail-party problem revisited: early processing and selection of multi-talker speech". In: *Attention, Perception, & Psychophysics* 77.5, pp. 1465–1487.

Cherry, E Colin (1953). "Some experiments on the recognition of speech, with one and with two ears". In: *The Journal of the acoustical society of America* 25.5, pp. 975–979.

Deleforge, Antoine, Florence Forbes, and Radu Horaud (2015). "Acoustic space learning for sound-source separation and localization on binaural manifolds". In: *International journal of neural systems* 25.01, p. 1440003.

Drullman, Rob and Adelbert W Bronkhorst (2000). "Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation". In: *The Journal of the Acoustical Society of America* 107.4, pp. 2224–2235.

Gardner, William G and Keith D Martin (1995). "HRTF measurements of a KEMAR". In: *The Journal of the Acoustical Society of America* 97.6, pp. 3907–3908.

Hawley, Monica L, Ruth Y Litovsky, and John F Culling (2004). "The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer". In: *The Journal of the Acoustical Society of America* 115.2, pp. 833–843.

Haykin, Simon and Zhe Chen (2005). "The cocktail party problem". In: *Neural computation* 17.9, pp. 1875–1902.

Hebrank, Jack and D Wright (1974). "Spectral cues used in the localization of sound sources on the median plane". In: *The Journal of the Acoustical Society of America* 56.6, pp. 1829–1834.

Hornstein, Jonas et al. (2006). "Sound localization for humanoid robots-building audio-motor maps based on the HRTF". In: *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE, pp. 1170–1176.

Keyrouz, Fakheredine and Klaus Diepold (2006). "An enhanced binaural 3D sound localization algorithm". In: *Signal Processing and Information Technology, 2006 IEEE International Symposium on*. IEEE, pp. 662–665.

Kim, Ui-Hyun, Kazuhiro Nakadai, and Hiroshi G Okuno (2015). "Improved sound source localization in horizontal plane for binaural robot audition". In: *Applied Intelligence* 42.1, pp. 63–74.

Langendijk, Erno HA and Adelbert W Bronkhorst (2002). "Contribution of spectral cues to human sound localization". In: *The Journal of the Acoustical Society of America* 112.4, pp. 1583–1596.

Roman, Nicoleta and DeLiang Wang (2008). "Binaural tracking of multiple moving sources". In: *IEEE Transactions on Audio, Speech, and Language Processing* 16.4, pp. 728–739.

Wenzel, Elizabeth M et al. (1993). "Localization using nonindividualized head-related transfer functions". In: *The Journal of the Acoustical Society of America* 94.1, pp. 111–123.

Windelspecht, M. and D. Sylvia S. Mader (2016). *Inquiry into Life*. McGraw-Hill Education. ISBN: 9781259426162. URL: `https://books.google.se/books?id=MF4cjwEACAAJ`.

5 Appendix

I received two reviews that were wildly different, but I tried to address certain issues presented.

Relevance of learning outcomes Here I received a 6/6 and a 4/6. I fixed the issues raised by adding section 3 on implementation and evaluation

Literature study Here I received a 1/6 and a 6/6. The reviews were highly conflicting, one saying the literature study was extensive but the other saying that the citations were inconsistent. I tried to fix all inconsistencies to my fullest ability.

Novelty/Originality Here I received a 0/6 and a 1/6. The literature review is not novel nor original and I guess it should not be.

Correctness Here I received a 4.5/6 and a 5/6. One reviewer said the literature review followed a logical path, and the other wanted more related work presented.

Clarity of presentation Here I received a 5/6 and a 4/6 and fixed a few structural elements of the report.