Improvement of Speech Intelligibility of Mobile Devices in Noisy Environments

Abstract

Multi-channel signal processing has gained popularity due to its practical application (e.g., speech recognition, noise reduction etc.) in a wide range of fields (i.e., military applications, music editing, audio information retrieval, etc.), and to the increased use of mobile and assistive communication devices. Without a doubt, increases in the computational power of these devices have allowed for the use of a wide range of techniques for noise reduction and to increase speech intelligibility. However, many of these are ineffective in many situations and, if not adjusted correctly for the particular environment, often introduce artifacts and distortions to the signal. Furthermore, these devices are often expensive, removing there availability from the general public, and proprietary, making them difficult to replicate and verify. Hence, there is a need for techniques for improving speech intelligibility in noisy environments which are open source, verifiable, and repeatable.

Objectives and Importance

The overall objective is to create a technique for improving speech intelligibility in noisy environments. To achieve this, the following subtopics must be addressed.

Sound Source Separation

Source separation problems in Digital Signal Processing (DSP) are those in which several signals have been mixed together into a combined signal and the objective is to recover the original component signals from the combined signal [1], [2]. The classical example of a source separation problem is the cocktail party problem, where a number of people are talking simultaneously in a room (e.g., at a cocktail party), and a listener is trying to follow one of the discussions. The human brain can handle this sort of auditory source separation problem, but it is a difficult problem in DSP. Several approaches have been proposed for the solution of this problem but development is currently still very much in progress. Some of the more successful approaches are Principal Component Analysis (PCA) and Independent Components Analysis (ICA), which work well when there are no delays or echoes present (i.e., the problem is simplified a great deal). The field of computational auditory scene analysis attempts to achieve auditory source separation using an approach that is based on human hearing. The human brain must also solve this problem in real time. In human perception, this ability is commonly referred to as Auditory Scene Analysis (ASA) or the cocktail party effect. These techniques will be investigated for separating an audio mixture and then Head-Related Transfer Functions (HRTFs) will be used to position the sources in their respective positions in 3D space.

Noise Reduction

Noise reduction is the process of removing “noise” from a signal. Noise is a relative term which, for this research, means any signal that is not of interest (i.e., non-speech signals). These can come from isotropic (i.e., coming from a human) and anisotropic (i.e., not coming from a human) sources which can be handled to some extent by the sound source separation technique. However, since these techniques generally rely on statistical processing, they often times have difficulty discerning between noise and signal(s) of interest if they are highly correlated. Software based solutions (i.e., spectral subtraction) and hardware based solutions (i.e., beamforming microphone arrays) will be explored.

Sound Source Localization

For the re-synthesis process, the 3D spatial locations of the isotropic sources must be determined. Once this is known, 3D binaural spatialization can be applied to the separated isotropic sources to position them in their appropriate locations in the 3D virtual auditory scene prior to its output. Hence, an effective method of sound source localization needs to be identified or develop. This will then be used to identify the 3D spatial locations of the isotropic sources and then this information will be used in the re-synthesis process.

Audio signals are usually recorded as a mixture of several sound sources and extracting knowledge about the sources spatial position is a difficult problem. However, there are many applications, such as video conferencing, surveillance, or source separation, that would benefit from this capability [3], [4]. Sound source localization is the task of locating a sound source(s) given a set of recorded mixtures. Using these mixtures, it is indirectly possible to obtain a source direction. There are many 3D sound localization methods that are used for various applications:

- Different types of sensor structures can be used such as microphone array and binaural hearing robot head [5]
- Optimal techniques such as neural networks, maximum likelihood, and Multiple Signal Classification (MUSIC)
- Real-time methods such as TDOA

Methods such as these, as well as novel approaches, will be investigated and/or developed, and used in the re-synthesis of the auditory scene.

Spatialization and Re-Synthesis using HRTFs

At this point, sound source separation and localization must be complete. The location of the identified isotropic sources will be used to re-synthesize the auditory scene. Initially, "generic" HRTFs (e.g., MIT's measurements from a KEMAR head microphone [6] or the CIPIC database [7])) will be used to spatialize the isotropic sound sources. However, it is known that generic HRTFs result in localization errors. It was shown that attempting to use HRTFs measured from a subject for the purpose of spatializing sounds, which is to be presented to a different subject, will result in a reduced fidelity of the perceived spatialization [8]. In consequence, the audio spatialization research community has pursued approaches that would "customize" a set of HRTF pairs to make it a better match for a subject who was not originally involved in their measurement. Several approaches have been developed that attempt to generate HRTFs without requiring inconvenient measurements while still achieving comparable performance to measured HRTFs including numerical computational methods, database interpolation, physical models, structural models and statistical models (for more information on customization models and methods see [9], [10]). Existing and novel models of generating customized HRTFs will be explored to reduce sound source localization errors.

HRTF Measurement and Listening Tests

The individualized measurements will be obtained using an HRTF measurement system such as the AuSIM HeadZap [11]. These types of systems measure impulse responses for both the left and right ears. Golay Codes [12] or Maximum Length Sequences (MLS) [6] will be used to generate a broad-spectrum stimulus signal delivered through a speaker(s). The responses for the normal hearing participants will be measured using miniature blocked-meatus microphones which are inserted into the ear canal. Under control of the system, the excitation sound is issued and both the left and right ear Head-Related Impulse Responses (HRIRs), which are the time-domain representation of the HRTFs, are captured. The system then provides these measured HRIRs as a pair of minimum-phase vectors and an additional delay value that represents the ITD.

In order to test the HRIRs localization performance, a GUI will be created in Matlab. A trial in this listening test was conducted in the following steps. First, a white Gaussian noise signal (Figure 1), which will be referred to as the input signal, was generated containing real white Gaussian noise of power 0 dBW. During a pilot study, it was determined that an appropriate sound level for the input signal is 72 dB which was measured using a Radioshack sound level meter (model #33-2055). Additionally, the duration of the input signal was 300 ms. Therefore, the test parameters conformed to recommendations for listening tests studies that indicate, for the best localization performance of the subjects, that the input signal should be between 40-80 dB and have a duration greater than 100 ms [13]. The second step is to convolve the input signal with the current trial's pair (left and right) of HRIRs in the time-domain to obtain a stereo signal. Finally, the resulting stereo signal is played to the subject and, using the computer mouse, he/she is required to point and click on the GUI to indicate where he/she estimates is the location of the emulated sound source.

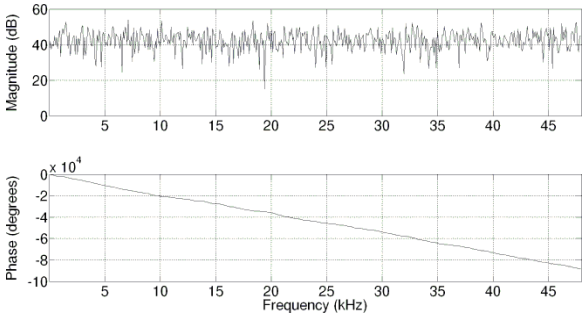

Figure 1: Example frequency and phase characteristics of white Gaussian noise.

Levels of Contribution

The main contribution of this independent study will be the creation of a device that transforms music into a multi-sensory experience. This simple assistive device can be used by individuals with hearing impairments to experience music through touch. If successful, the materials will be disseminated at relevant conferences and/or journals. This work can then be used by students interested in the areas of DSP, real-time audio processing, and assistive technology.

Outcomes

At the end of the semester, students should have an understanding of:

- Coding Matlab
- Audio Signal Processing
- Blind Source Separation (BSS) techniques
- Audio source localization techniques
- Spatializing audio using HRTFs
- Audio de-noising techniques
- Creating and conduction listening test

Assessment of Activities and Deliverables

The activities described above will be assessed on a continual basis through progress reports and weekly meetings. The direction of the activities will be refined based on this information. As shown in the schedule below, the student will periodically provide progress reports and their current code, and the instructor will evaluate the student's progress based on this.

Tentative Schedule

Week	Topics	Assignment
1	Literature Review	Literature Review Report
2	Sound Source Separation	Progress Report and Code
3	Noise Reduction	Progress Report and Code
4	Sound Source Localization	Progress Report and Code
5	Spatialization and Re-Synthesis	Progress Report and Code
6	HRTF Measurement	Progress Report and Code
7	Listening Tests	Progress Report
8	Presentation and Final Report Preparation	Final Report and Final Code

Grading

Plus and minus grading will be used when determining final grades. Final grades are computed by first finding the average score in each category described in the table below on the right. All scores are normalized to a scale of 0 to 100 before being averaged. The average score for each category is then used to compute the weighted average according to the weights in the second table below.

Letter Grade	*% of Total Points*
A+	96% & Above
A	93% – 95.99%
A-	88% – 92.99%
B+	85% – 87.99%
B	82% – 84.99%
B-	78% – 81.99%
C+	75% – 77.99%
C	72% – 74.99%
C-	68% – 71.99%
D+	65% – 67.99%
D	62% – 64.99%
D-	58% – 61.99%
F	Less than 58%

Category	*% of Final Grade*
Reports	30%
Code	30%
Presentation	20%
Final Report	20%

References

[1] A. O. V. Bimbot, "A General Flexible Framework for the Handling of Prior Information in Audio Source Separation," 2010.

[2] Y. Salaün *et al.*, "The Flexible Audio Source Separation Toolbox Version 2.0," in *ICASSP*, 2014.

[3] M. Brandstein and D. Ward, *Microphone Arrays: Signal Processing Techniques and Applications*. Springer Science & Business Media, 2013.

[4] S. Makino, T.-W. Lee, and H. Sawada, *Blind Speech Separation*. Springer, 2007.

[5] H. Nakashima and T. Mukai, "3D Sound Source Localization System Based on Learning of Binaural Hearing," in *2005 IEEE International Conference on Systems, Man and Cybernetics*, 2005, vol. 4, pp. 3534–3539.

[6] B. Gardner and K. Martin, “HRTF Measurements of a KEMAR Dummy-Head Microphone,” *Massachusetts Institute of Technology (MIT) Media Laboratory Vision and Modeling Group*. [Online]. Available: http://sound.media.mit.edu/resources/KEMAR.html. [Accessed: 08-Mar-2014].

[7] V. Algazi, R. Duda, D. Thompson, and C. Avendano, “The Cipic HRTF Database,” *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2001. [Online]. Available: http://interface.cipic.ucdavis.edu/. [Accessed: 08-Mar-2014].

[8] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman, “Localization Using Nonindividualized Head-Related Transfer-Functions,” *J. Acoust. Soc. Am.*, vol. 94, pp. 111–123, 1993.

[9] J. Blauert, *Spatial Hearing : The Psychophysics of Human Sound Source Localization, 2nd ed*. Cambridge, MA: MIT Press, 1997.

[10] P. Satarzadeh, V. Algazi, and R. Duda, “Physical and Filter Pinna Models Based on Anthropometry,” in *122nd Convention of the Audio Engineering Society (AES)*, 2007.

[11] AuSIM Inc., “Corporate Website,” 2012. [Online]. Available: http://www.ausim3d.com/. [Accessed: 08-Mar-2014].

[12] S. Foster, “Impulse Response Measurement Using Golay Codes,” *IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 11, pp. 929–932, 1986.

[13] G. Wersényi, “Localization in a HRTF-Based Minimum-Audible-Angle Listening Test for GUIB Applications,” *Electron. J. - Tech. Acoust.*, vol. 1, pp. 1–16, 2007.