ICASSP, Brisbane, Australia, 19-24, April, 2015

MULTI-SHIFT PRINCIPAL COMPONENT ANALYSIS BASED PRIMARY COMPONENT EXTRACTION FOR SPATIAL AUDIO REPRODUCTION

Jianjun He, and *Woon-Seng Gan*

Digital Signal Processing Lab, School of Electrical and Electronic Engineering,
Nanyang Technological University, Singapore
{jhe007@e.ntu.edu.sg, ewsgan@ntu.edu.sg}

ABSTRACT

In spatial audio analysis-synthesis, one of the key issues is to decompose a signal into primary and ambient components based on their spatial features. Principal component analysis (PCA) has been widely employed in primary component extraction, and shifted PCA (SPCA) is employed to enhance the primary extraction for input signals involving the inter-channel time difference. However, SPCA generally requires the primary components to come from one direction and cannot produce good results in the case of multiple directions. To solve this problem, we propose multi-shift PCA (MSPCA) by extending SPCA to multiple shifts. Two structures of MSPCA with different weighting methods are discussed. From the results of our simulations and listening tests, the proposed consecutive MSPCA with proper weighting is found to be superior to the conventional PCA and SPCA based primary extraction methods.

Index Terms**—**principal component analysis (PCA), primary-ambient extraction (PAE), spatial audio, multiple sources, time shifting

1. INTRODUCTION

Increasing prevalence of 3D video technology calls for a more immersive listening experience to better match 3D visual effects, resulting in a growing need for 3D audio or spatial audio reproduction. Primary component that is directional and ambient component that is diffuse are two critical elements in spatial audio processing. Different processing schemes should be independently applied to the primary and ambient components to enhance the perception of spatial audio [1]. However, the primary and ambient components are not stored separately in conventional audio formats including stereo and 5.1, which necessitates primary-ambient extraction (PAE). In recent years, PAE has been widely applied in spatial audio processing [2], [3], audio mixing [4]-[6], spatial audio coding [7], [8], immersive 3D sound system [9]-[11], and natural sound rendering headphone systems [12].

To date, many approaches have been proposed for PAE, including time-frequency masking [13], least-squares [14], mixing model classification [15], estimation using multi-channel pair wise correlations [16], ambient phase estimation [17], and principal component analysis (PCA) [1], [18]-[22]. PCA remains one of the most widely studied methods in PAE. Applied in stereo signals, PCA transforms the signal space into two orthogonal basis vectors using the Karhunen-Loève transform [23]. As the primary component is usually stronger than the ambient component, the vector corresponds to the larger eigenvalue is used as a projection basis for the primary components.

When the primary component is not completely correlated at zero lag, the performance of PAE is severely degraded. The performance degradation includes inaccurate estimation of inter-channel time difference (ICTD) and inter-channel level difference (ICLD) of the primary components, which results in erroneous sound localization. To solve this problem, shifted PCA (SPCA) is introduced [21] to shift the input signal according to the estimated ICTD prior to PCA. However, one single shift only accounts for one direction, which is improper for the primary components that consist of sound sources from multiple directions. Thus, a common approach is to decompose the signal into subband before the extraction, assuming that only one source is dominant in each subband [14], [24]. On this note, the directions of multiple sources can be tracked [25] and localized [26] in the presence of ambient noise. Nevertheless, subband PAE approaches become problematic when the spectra of the sources in the primary components overlap in certain subbands. Meanwhile, timbre change is an inevitable problem in subband PAE.

In this paper, we investigate the primary component extraction (or primary extraction for short) with multiple directions by extending the single shift SPCA to multiple shifts. These shifts are performed based on the ICTD estimation. While in the output, the extracted primary components are correspondingly shifted back, weighted and summed to obtain the final result of the extracted primary components. We refer to the proposed method as multi-shift PCA (MSPCA) in this paper. The typical structure of MSPCA is shown in Fig. 1.

The rest of the paper is organized as follows. In Section 2, we review the stereo signal model, and PCA, SPCA based

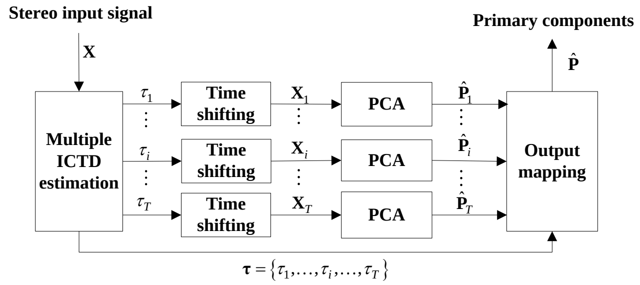

Fig. 1 Typical structure of MSPCA (MSPCA-T). Stereo input signal $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1\}$; τ_i is the *i*th estimated ICTD (T is the total number of ICTDs); $\mathbf{X}_i$ and $\hat{\mathbf{P}}_i$ are the corresponding shifted signal and extracted primary component, respectively. The final output of the extracted primary components is denoted by $\hat{\mathbf{P}}$. In the case of consecutive MSPCA, the time shifting is applied for every individual lag.

primary extraction methods. Section 3 discusses the proposed MSPCA based primary extraction. Section 4 presents a series of performance comparisons, including simulation results and subjective listening tests. Finally, we conclude this work in Section 5.

2. PCA AND SPCA BASED PRIMARY EXTRACTION

In this section, we introduce the stereo signal model and explain its key assumptions. Based on this model, PCA and SPCA are employed in primary extraction.

2.1. Stereo Signal Model

In general, we consider a stereo signal to consist of two components: (i) a directional component referred to as the primary component; and (ii) a diffused component referred to as the ambient component [1]. Denoting one frame of the time-domain stereo signals as $\mathbf{x}_0, \mathbf{x}_1$, we formulate the basic signal model as:

$$\mathbf{x}_0 = \mathbf{p}_0 + \mathbf{a}_0,\ \mathbf{x}_1 = \mathbf{p}_1 + \mathbf{a}_1, \qquad (1)$$

where $\mathbf{p}_0$, $\mathbf{p}_1$ and $\mathbf{a}_0$, $\mathbf{a}_1$ are the primary and ambient components in the two channels of the stereo signal, respectively.

The statistical characterization for the primary and ambient components in this model is that the primary components are assumed to be correlated while the ambient components are uncorrelated. The unit correlation of the primary components is usually realized by amplitude panning between two channels, i.e., $\mathbf{p}_1 = k\mathbf{p}_0$, where k is the primary panning factor [1]. Higher values of k indicate that the primary component is panned more towards channel 1. Furthermore, the primary and ambient components are assumed to be uncorrelated with each other. The diffuseness of the ambient component leads to a balance of ambient power between the two channels. To determine the power difference between the primary and ambient components, we introduce the primary power ratio, which is the percentage of the primary component power in the input signal.

2.2. Primary Extraction using PCA and SPCA

Based on the signal model, primary extraction using PCA can be obtained by eigenvalue decomposition of the input covariance matrix. The results of extracted primary components using PCA can be obtained as [22]

$$\hat{\mathbf{p}}_0 = \frac{\mathbf{x}_0 + k\mathbf{x}_1}{1+k^2},\ \hat{\mathbf{p}}_1 = \frac{k\mathbf{x}_0 + k^2\mathbf{x}_1}{1+k^2}. \qquad (2)$$

However, it is unlikely for any stereo input signals to satisfy all the assumptions of the stereo signal model. As discussed in [27], correlated signals can be amplitude panned as well as time shifted. To overcome the limitation of PCA and improve the extraction of time shifted primary components, SPCA is proposed [21]. In SPCA, the stereo input signals are first time-shifted according to the estimated ICTD of the primary component before PCA. Subsequently, the extracted primary components are shifted back using the same ICTD. The critical step of SPCA is to estimate the correct ICTD for the shifting. There has been extensive research on ICTD estimation (see [28]-[30] and references therein). Based on Jeffress coincidence model [28], the inter-channel cross-correlation coefficient (ICC) of different time lags is calculated and the lag number that corresponds to the maximum of ICC is the estimated ICTD of the primary components in the stereo signal.

3. MSPCA BASED PRIMARY EXTRACTION

In many applications of spatial audio, concurrent sound sources from different directions and even the reflections of these sound sources (image sources) are frequently encountered in the stereo mix. These directions of the sources and reflections imply multiple different ICTDs. In such cases, SPCA with one single shift that corresponds to one single direction becomes problematic. Therefore, to account for multiple directions in the primary components of the stereo signal, we extend SPCA from one single shift to multiple shifts, and develop MSPCA for primary extraction. The typical structure of the MSPCA (MSPCA-T) is shown in Fig. 1. First, several ICTDs are estimated from the stereo input signal by finding the peaks in the short time cross correlation function [31]. Next, the input signal is time shifted according to the estimated ICTDs [21]. For every shifted version, PCA is applied to obtain the extracted primary components. Finally, the extracted primary components of all shifted versions are properly mapped, weighted and linearly summed to obtain the final output of the extracted primary components. Note that the weights are computed according to the significance of each shifted version.

Combining the selective time shifting with the significance based weighting method, a consecutive structure for MSPCA can also be employed. Instead of shifting the input signal according to a few selected ICTDs, we perform the shifting consecutively lag by lag. Subsequently, PCA based primary extraction is employed for each shifted version. Before reversing the one-lag shifting and adding to the final output, the extracted primary components of each shifted version are weighted based on the significance of each shifted version. By assuming that those shifted versions having higher ICC are more significant, the weights are set higher for the shifted version with higher ICC. Via this ICC based weighting method, we can unify the consecutive MSPCA and MSPCA-T.

Let the stereo input signal be $\mathbf{X} = \{\mathbf{x}_0, \mathbf{x}_1\}$. The shifted signal is $\mathbf{X}_l = \{\mathbf{x}_0, \vec{\mathbf{x}}_1^l\}$ with nth sample of $\vec{\mathbf{x}}_1^l$ shifted by l lags, as $\vec{x}_1^l(n) = x_1(n-l)$, where $l \in [-L, L]$. The extracted primary components at the lth shifted version $\hat{\mathbf{P}}_l$ are computed using PCA. The final output of the extracted primary components $\hat{\mathbf{P}}$ can be expressed as a weighted sum of the shifted back version of $\hat{\mathbf{P}}_l$. The nth sample of $\hat{\mathbf{P}}$ (either $\hat{\mathbf{p}}_0$ or $\hat{\mathbf{p}}_1$) is hence obtained by

$$\hat{P}(n) = \sum_{l=-L}^{L} w_l \hat{P}_l(n+l), \tag{3}$$

where $w_l \geq 0$ is the weight applied on $\hat{\mathbf{P}}_l$. To retain the overall signal power, the weights shall sum up to one, i.e., $\sum_{l=-L}^{L} w_l = 1$. Since the weights in consecutive MSPCA are proportional to the ICC of each lag, a straightforward way to obtain the weights is to employ the exponent of the ICC, i.e., $w_l = \phi_l^a \Big/ \sum_{l=-L}^{L} \phi_l^a$, where a is the exponent and ϕ_l is the ICC of lag l. Larger values of a leads to sparser weights. Examples of the exponent selection for the weighting methods are shown in the following section.

4. EXPERIMENTS AND DISCUSSIONS

To evaluate the performance of the proposed MSPCA based primary extraction, a number of simulations and subjective listening tests are conducted. In our experiments, primary components consist of a speech signal and a music signal, which are amplitude panned by a factor of three and time shifted by 20 lags, towards the channel 1 and channel 0, respectively; and uncorrelated white Gaussian noise is used as the ambient component. Subsequently, the primary and ambient components are linearly mixed by setting the root-mean-square power of the speech, music and ambient component to be equal, which means primary power ratio equals to 0.67. Next, PCA, SPCA and MSPCA with different settings are employed to extract primary components from the synthesized stereo signals. The searching range for ICTD is ±50 lags, which is around 2ms for sampling frequency at 44.1 kHz. Finally, the performance of primary extraction using these approaches is compared using objective metrics and subjective testing.

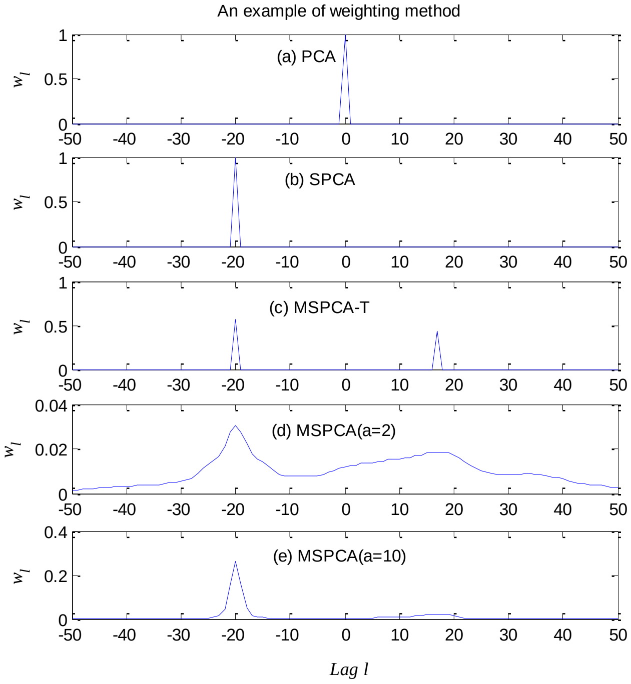

Fig. 2 An illustration of the weighting methods in PCA, SPCA and MSPCAs. Negative and positive lags correspond to the direction towards the channel 1 and channel 0, respectively.

It can be found that PCA and SPCA can be considered as special cases of MSPCA by specifically setting the weights. Both PCA and SPCA have only one nonzero weights, but at different lags. While the corresponding lag for the unit weight in PCA is always zero, SPCA places the unit weight at the lag corresponding to maximum ICC. Since all weights shall sum up to one, this maximum weight for PCA and SPCA will be exactly equal to one. MSPCA-T can detect the two ICTDs by peak finding. After normalization, we can consider it having two nonzero weights at the two corresponding lags. For consecutive MSPCA, we examine two exponent values, namely, a = 2, 10. Summarizing all different settings for these approaches, the weighting methods are compared in Fig. 2. As discussed, PCA and SPCA have only one nonzero weight at zero lag and -20 lag, respectively. For MSPCA-T, two weights are applied at two distinct lag positions, though the positive ICTD for the music is not as accurate as the negative ICTD for the speech. For consecutive MSPCA with different exponent values, the non-zero weights are found for all the lags, and apparently higher weights are given to those lags that are closer to the directions of the primary components. As the exponent value a increases, the differences among the weights at various lags become more significant. When a is high (e.g., a=10),

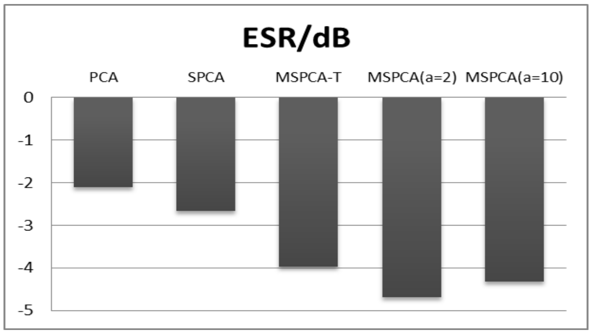

Fig. 3 Objective performance on extraction accuracy measured by ESR for PCA, SPCA, MSPCAs.

the weighting method in consecutive MSPCA becomes similar to SPCA, as seen from Fig. 2(b) and Fig. 2(e).

After applying these approaches, the objective performance on the extraction accuracy of the primary component is determined by error-to-signal ratio (ESR), which can be computed by [23]

$$\mathrm{ESR(dB)} = 10\log_{10}\left[\left(\frac{\left\|\hat{\mathbf{p}}_0 - \mathbf{p}_0\right\|_2^2}{\left\|\mathbf{p}_0\right\|_2^2} + \frac{\left\|\hat{\mathbf{p}}_1 - \mathbf{p}_1\right\|_2^2}{\left\|\mathbf{p}_1\right\|_2^2}\right)\Big/ 2\right]. \quad (4)$$

A better performance is achieved when ESR is smaller. The ESR results for these approaches are illustrated in Fig. 3. It is obvious that MSPCAs generally perform better than PCA or SPCA by having smaller ESR. It is also quite interesting to observe that consecutive MSPCA approaches outperform MSPCA-T. On this note, the accuracy in the estimation of the number of the directions and the associated ICTDs are extremely critical for MSPCA-T. Failure to accurately estimate any ICTDs will degrade the overall extraction performance, as observed here. By contrast, consecutive MSPCA mitigates this problem by applying weights at all lags. Furthermore, the averaging of the ambient components across various shifted versions could also reduce ambient leakage in the extracted primary components. Between the two consecutive MSPCA approaches, MSPCA(a=2) performs better than MSPCA(a=10). Therefore, the exponent applied on the ICC for the weights in consecutive MSPCA cannot be too large.

In addition to the objective assessment on the error performance, subjective testing of localization accuracy of the primary extraction was also conducted. The testing method was based on MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [32], [33]. Nine signals, including primary components extracted using the five methods, one known reference, one hidden reference and two anchors, were tested. The subjects were asked to rate a score of 0-10, where a score of 0 denotes the worst localization (i.e., the two directions are reversed), and a score of 10 denotes the same directions perceived as the reference. When at least one direction is accurate, a score of no less than 5 shall be given, and a score of 3-7 shall be appropriate for those signals with perceived directions neither too close nor too bad. Finally, 12 subjects participated in the experiment and the results are shown in Fig. 4. Generally, MSPCAs produce more accurate localization of the primary components among these testing methods. Similar to the observation in ESR, MSPCA(a=2) performs the best and MSPCA(a=10) degrades the localization significantly. Therefore, it can be concluded that consecutive MSPCA with proper weighting can help improve both the extraction accuracy and localization accuracy of the primary components when there are multiple directions.

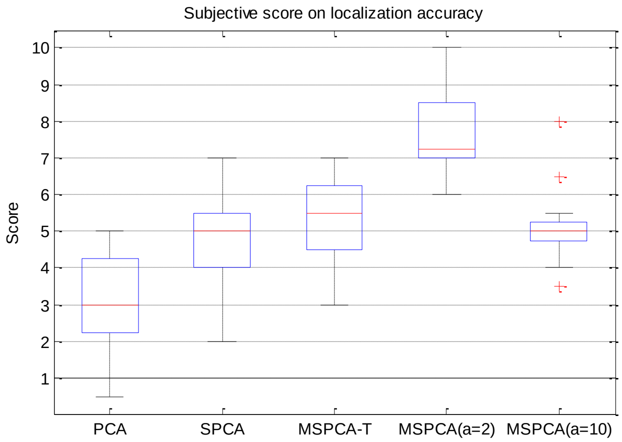

Fig. 4 Subjective performance on localization accuracy for PCA, SPCA, MSPCAs.

5. CONCLUSIONS

In this paper, we investigated the problem of primary component extraction from stereo signals that consist of primary components coming from multiple concurrent directions. To account for these directions, a multi-shift PCA approach is proposed in this paper. Two different structures of MSPCA are examined. While MSPCA with typical structure is simpler, its performance relies heavily on the correct estimation of the ICTDs. By contrast, consecutive MSPCA is more robust by applying weights on all shifted versions. The weighting method for different shifted versions is found to be critical to the extraction performance. In general, applying the exponential function of ICC with proper exponent value as the weightings yields a good performance in terms of the extraction accuracy as well as localization accuracy. Future works include study on how to determine the exponent value in the ICC based weighting method as well as other weighting methods.

ACKNOWLEDGMENT

This work is supported by the Singapore Ministry of Education Academic Research Fund Tier-2, under research grant MOE2010-T2-2-040.

REFERENCES

[1] M. M. Goodwin and J. M. Jot, “Primary-ambient signal decomposition and vector-based localization for spatial audio coding and enhancement,” in *Proc. ICASSP*, Hawaii, 2007, pp.9-12.

[2] F. Rumsey, *Spatial Audio*. Oxford, UK: Focal Press, 2001.

[3] J. Breebaart and C. Faller, *Spatial audio processing: MPEG surround and other applications*. Chichester, UK: John Wiley & Sons, 2007.

[4] M. R. Bai and G. Y. Shih, “Upmixing and downmixing two-channel stereo audio for consumer electronics,” *IEEE Trans. Consum. Electron.*, vol. 53, no. 3, pp. 1011-1019, Aug. 2007.

[5] S. Y. Park, S. Lee, and D. Youn, “Robust representation of spatial sound in stereo-to-multichannel upmix,” in *Proc. 128th Audio Eng. Soc. Conv.*, London, UK, 2010.

[6] C. Faller and J. Breebaart, “Binaural reproduction of stereo signals using upmixing and diffuse rendering,” in *Proc. 131th Audio Eng. Soc. Conv.*, New York, 2011.

[7] C. Faller and F. Baumgarte, “Binaural cue coding-part II: schemes and applications,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp.520-531, Nov. 2003.

[8] M. M. Goodwin and J. M. Jot, “Binaural 3-D audio rendering based on spatial audio scene coding,” in *Proc. 123rd Audio Eng. Soc. Conv.*, New York, 2007.

[9] W. S. Gan, E. L. Tan, and S. M. Kuo, “Audio projection: directional sound and its application in immersive communication,” *IEEE Signal Processing Mag.*, vol. 28, no. 1, pp. 43-57, Jan. 2011.

[10] E. L. Tan, and W. S. Gan, “Reproduction of immersive sound using directional and conventional loudspeakers,” *J. Acoust. Soc. Amer.*, vol. 131, no. 4, pp. 3215-3215, Apr. 2012.

[11] E. L. Tan, W. S. Gan, and C. H. Chen, “Spatial sound reproduction using conventional and parametric loudspeakers,” in *Proc. APSIPA ASC*, Hollywood, CA, 2012.

[12] K. Sunder, J. He, E. L. Tan, and W. S. Gan, “Natural sound rendering for headphones,” in press, *IEEE Signal Processing Magazine*, DOI: 10.1109/MSP.2014.2372062, Mar. 2015.

[13] C. Avendano and J. M. Jot, “A frequency-domain approach to multichannel upmix,” *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 740-749, Jul./Aug. 2004.

[14] C. Faller, “Multiple-loudspeaker playback of stereo signals,” *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051-1064, Nov. 2006.

[15] A. Härmä, “Classification of time-frequency regions in stereo audio,” *J. Audio Eng. Soc.*, vol. 59, no. 10, pp. 707-720, Oct. 2011.

[16] J. Thompson, B. Smith, A. Warner, and J. M. Jot, “Direct-diffuse decomposition of multichannel signals using a system of pair-wise correlations,” in *Proc. 133rd Audio Eng. Soc. Conv.*, San Francisco, 2012.

[17] J. He, W. S. Gan, and E. L. Tan, “Primary-ambient extraction using ambient phase estimation with a sparsity constraint,” *IEEE Signal Process. Letters*, vol. 22, no. 8, pp. 1127-1131, Aug. 2015.

[18] J. Merimaa, M. M. Goodwin, and J. M. Jot, “Correlation-based ambience extraction from stereo recordings,” in *Proc. 123rd Audio Eng. Soc. Conv.*, New York, 2007.

[19] R. Irwan and R. M. Aarts, “Two-to-five channel sound processing,” *J. Audio Eng. Soc.*, vol. 50, no. 11, pp. 914-926, Nov. 2002.

[20] Y. H. Baek, S. W. Jeon, Y. C. Park, and S. Lee, “Efficient primary-ambient decomposition algorithm for audio upmix,” in *Proc. 133rd Audio Eng. Soc. Conv.*, San Francisco, 2012.

[21] J. He, E. L. Tan, and W. S. Gan, “Time-shifted principal component analysis based cue extraction for stereo audio signals,” in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 266-270.

[22] J. He, E. L. Tan and W. S. Gan, “Linear estimation based primary-ambient extraction for stereo audio signals,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 2, pp. 505-517, Feb. 2014.

[23] I. Jolliffe, *Principal component analysis, 2nd ed.*, New York: Springer-Verlag, 2002.

[24] J. He, W. S. Gan and E. L. Tan, “A study on the frequency-domain primary-ambient extraction for stereo audio signals,” in *Proc. ICASSP*, Florence, Italy, 2014, pp. 2868-2872.

[25] N. Roman and D. L. Wang, “Binaural tracking of multiple moving sources,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 728-739, May 2008.

[26] J. Woodruff and D. L. Wang, “Binaural localization of multiple sources in reverberant and noisy environments,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 5, pp. 1503-1512, Jul. 2012.

[27] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. Cambridge, MA: MIT Press, 1997.

[28] A. Jeffress, “A place theory of sound localization,” *Journal of Comparative and Physiological Psychology*, vol. 41, no. 1, pp. 35-39, Feb. 1948.

[29] W. A. Yost, “Perceptual models for auditory localization,” in *Proc. 12th Audio Eng. Soc. Int. Conf.*, Copenhagen, Denmark, 1993.

[30] P. X. Joris, P. H. Smith, and T. Yin, “Coincidence detection in the auditory system: 50 years after Jeffress,” *Neuron*, vol. 21, no. 6, pp.1235-1238, Dec. 1998.

[31] The MathWorks, Inc. “Find the local maxima.” Internet: http://www.mathworks.com/help/signal/ref/findpeaks.html, [Apr. 10, 2013].

[32] ITU. Recommendation BS.1534-1: Method for the subjective assessment of intermediate quality levels of coding systems, Jan. 2003.

[33] E. Vincent, “MUSHRAM: A MATLAB interface for MUSHRA listening tests, 2005.” Internet: http://www.elec.qmul.ac.uk/people/emmanuelv/mushram/. [Apr. 05, 2013].