타악기 음원 분리에 기반한 모노–스테레오 업믹싱 기법

최근우
한국전자통신연구원
gnu@etri.re.kr

A Mono–To–Stereo Upmixing Algorithm Based on the Harmonic-Percussive Separation

Keunwoo Choi
Electronics and Telecommunications Research Institute (ETRI)

Abstract

In this research, a mono-to-stereo upmixing algorithm based on music source separation is proposed. For the upmixing, a harmonic and percussive separation for jazz music is implemented. Then, the sources are re-panned by equalizing the loudness of left and right sides of listeners in the one proposed approach. In the other approach, the harmonic sources are spread by a decorrelator while the percussive sources are panned to the center. In the experiments, the re-panning algorithm showed advanced performance in terms of localization and timbral quality.

1. Introduction

Before stereophonic recordings were introduced in 1950s, audio content was delivered in mono channel format. Still, music released in those years is consumed and played in mono channel. There have been some researches to improve the audio quality by upmixing them into stereo-channel signals.

For the monaural upmixing which remix mono channel signals into N-channel signals, several algorithms have been proposed. A simple method was proposed in [1], where the decorrelation technique is adopted to generate an artificial reverberation. This can be also used as an upmixing method, resulting in stereo or even multi-channel signals. Although some timbral degradation is observed, as will be reported in this research, this algorithm provides a very robust method as there is no assumption on the source signal. FitzGerald proposed a mono-to-stereo upmixing based on vocal, harmonic, and percussive source separation algorithms in [2], where the signals were separated using median-filtering [3], and tensor factorization [4]. In [2], however, the mixing was performed using digital audio workstation by people.

Sound source separation has been employed in many other upmixing algorithms for better localization of sound objects. Cobos et al., Kamado et al., and Shim et al. proposed multichannel upmixing algorithms using stereo mixtures and source separation algorithms in [5-7]. In general, those algorithms separate sound objects from original mixtures and relocate them at the estimated positions. The positions are automatically estimated during the separation process as the separation algorithms were based on position estimation and clustering of time-frequency bins.

We propose an algorithm for upmixing from mono audio signals to stereo signals using sound source separation. We mainly focus on the jazz music, as the popular songs released before 1950s are mostly jazz. Upmixing audio signals based on the sound source separation raise an issue of how to re-locate the audio sources. We propose a two different approach during mixing the separated sources. Both of them intend to clear localizations and wide source width, as well as preservations of the timbre. In the reported experiments, the performances of the proposed algorithms are assessed through subjective listening tests.

In the section 2, the source separation technique used in the proposed algorithms and the comparison algorithm are described. The proposed algorithms are introduced in the section 3. In the section 4, the procedure and results of the subjective tests are addressed, as well as discussing the results. Finally, we conclude the research in the section 5.

2. Background

2.1. Tonality-based Source Separation

In [8], a single–channel harmonic and percussive separation algorithm for jazz music was proposed. In the research, a Nonnegative Matrix Factorization [9] was adopted to decompose the spectrogram of the source signal. Tonality calculation in [10] was used to determine whether spectral bands in the decomposed signal as harmonic or percussive.

As the percussive part of jazz music mainly consists of noise-like sounds, the adoption of tonality measure resulted in advanced performance for the separation of jazz music. During the separation, the mono signals are separated into harmonic and percussive signals by time–frequency masking and inverse Short-Time Fourier Transforms (STFT). In the Figure 1, the block diagram is illustrated where $\mathbf{X}$ is the original source and $\mathbf{M_H}$ and $\mathbf{M_P}$ are masks for harmonic and percussive sources.

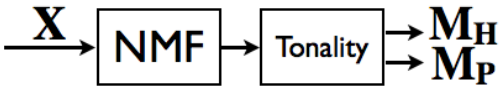

Figure 1. The block diagram of harmonic and percussive source separation in the STFT domain – The calculation of masks

2.2. Comparison algorithm

In [1], a decorrelation–based upmixing algorithm was introduced. It generates mutually decorrelated stereo signals from mono signals and feed them into the left and right channels.

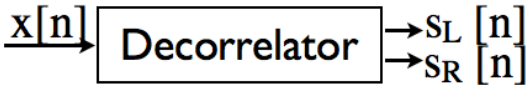

Figure 2. The block diagram of comparison algorithm

Decorrelation can be realized by various methods. In the proposed research, a filter-based approach in [11] is adopted with 1024 samples and an exponential window. Two different decorrelation filters are generated and applied to the mono signal.

This algorithm results in a randomly located stereo signal with a source width, controlled by the length of the filter. The complexity is not high and the performance is robust to the source signal. However, time-smearing due to the filtering results in the timbral degradation.

3. The Proposed Algorithms

In the proposed algorithms, the mono signal is separated into harmonic and percussive sources using the technique proposed in [8]. During the following stage, we propose two different approaches for mixing the separated sources, *Re-panning* and *selective decorrelation*, as in the following figures.

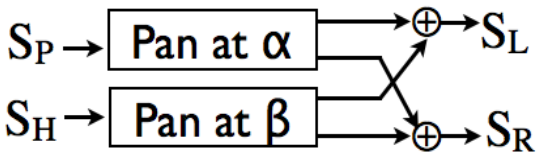

Figure 3. The block diagram of the proposed algorithm, the re–panning approach (in the STFT-domain).

Figure 3 shows the block diagram of the re-panning approach. In this approach, the separated sources are panned to the different side. They are located at the azimuth angels of α and β, which are determined under following two rules; i) maximizing the overall source width, and ii) equalizing the loudnesses of left– and right– channel signals. In the procedure, we calculate the loudness of each separated signal, compare them loudness, and pan one of the sources with less loudness to the end of one side, either left or right. Then we pan the other source to the other side at the angle which equalizes the loudness of left channel and right channel. The tangential law is used in the panning and described in Eq.(1), where θ_0 is the azimuth angle of loudspeakers, θ_T is the panning angle for the source, and g_1, g_2 are the gains for stereo loudspeakers.

$$\frac{\tan\theta_T}{\tan\theta_0} = \frac{g_1 - g_2}{g_1 + g_2} \tag{1}$$

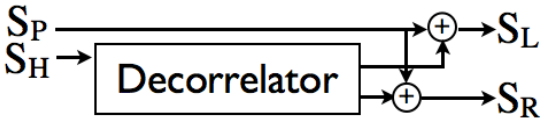

Figure 4. The block diagram of proposed algorithm, the selective decorrelation approach.

Figure 4 shows the block diagram of the selective decorrelation approach. In this approach, the harmonic sources are decorrelated to make the overall source width wide, while the percussive sources are panned to the center. Harmonic sources are chosen to be decorrelated as the smearing effects in the time-domain due to the decorrelation are more perceivable in impulsive sources.

4. Experiments

4.1. Overview

Subjective tests were conducted to verify the performance of the proposed algorithm by 17 participants. As an anchor, mono signal and 4 kHz low–pass filtered signal was used for spatial attributes and timbre, respectively. Genelec 8050A loudspeakers were used for the playback of the items. The loudness of items in different systems were equalized according to the ANSI Standard, [12].

4.2. Test Items

As MUSHRA test [13] requires a reference signal, the test excerpts should have stereophony mixtures though the proposed algorithms target monaural signals. As a result, 5 commercial jazz excerpts released in 1957 – 1959 were selected. The excerpts consist of piano, saxophone, trumpet, bass, and drums. They are downmixed by the addition of left and right channels for the generation of monaural signals. The bit depth and sampling rate of the excerpts are 16 and 44,100 Hz, respectively.

4.3. Attributes

Three attributes, Stereophonic Image Quality (SIQ), Ensemble Source Width (ESW), and timbre preservation, were selected. The SIQ is defined in [14] as *'how much the system is similar to the reference in terms of sound image locations and sensations of depth and reality of the audio event'*. In [15], the ESW is defined as '*the overall width of a defined group of sources*', in which the group is defined as all the sources in the stereo signals in the experiments.

In addition to those criteria, the Basic Audio Quality (BAQ) is calculated as a weighted sum of SIQ, ESW, and timbre. In [16], Rumsey et al. proposed Eq.(2) to calculate the overall quality regarding timbre, frontal localizations, and surround effects. As the 'Frontal' and 'Surround' are considered as a 'Spatial' attribute, a modified version of Eq.(2), is applied as Eq.(3).

$$BAQ = 0.80Timbre + 0.30Frontal + 0.09Surround - 18.7 \quad (2)$$

$$BAQ = 0.80Timbre + 0.39Spatial - 18.7 \quad (3)$$

The spatial qualities and timbral quality may have trade-off relationships. Therefore the optimization of each algorithm is not unique. Therefore, we empirically set the parameters of each algorithm to show similar source widths.

4.4. Results and Discussions

Figure 5 – Figure 8 show the results of subjective tests for each item with their means and confidential intervals. '*Stereo*', '*Sel. Decorr*', '*Decorr*', and '*Panning*' in the legends indicate the reference signals, proposed signals (selective decorrelated), comparison signals (decorrelated), and the other proposed signals (re-panned), respectively.

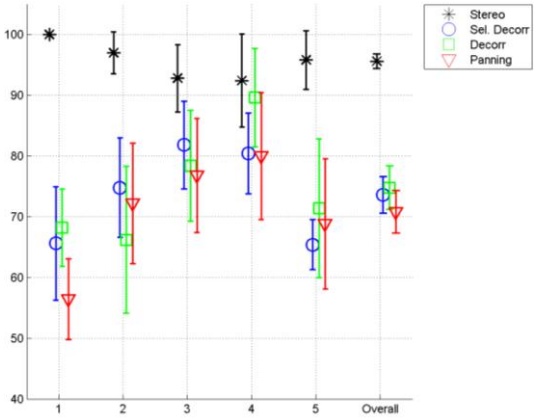

Figure 5. The results of subjective tests for ESW

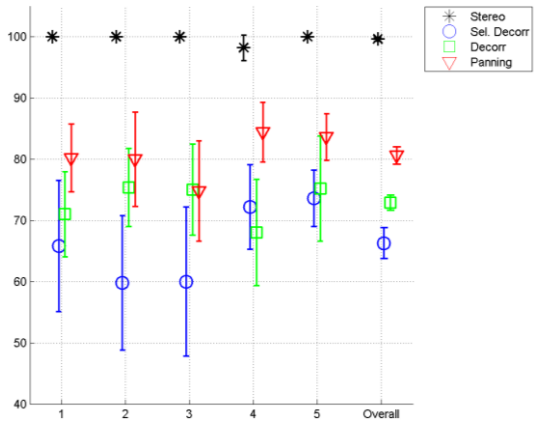

Figure 6. The results of subjective test for SIQ

The results of spatial qualities are plotted in Figure 5 and Figure 6. There are variances among items for ESW scores, but there is no significant difference in overall. We can interpret that the ESW of each algorithm are equalized well, as we intended. In the item 3 and 4, some participants gave the comparison signals higher score than the reference signals. Unlike MUSHRA tests for coding quality evaluation, such participants are not excluded, as the source width of some system can be wider than that of the reference signals.

In the Figure 6, SIQ scores are plotted. The re-panning approach showed advanced performance for this attribute by 8 points in overall. However, one of the proposed algorithms, Selective decorrelated signals obtained the lowest score. This is interesting as the selective decorrelation localize the separated percussive source, at least, to the center, while the comparison algorithm arbitrary localizes the all source by the decorrelation. For the similar ESW scores, the selective decorrelation excessively spread the separated harmonic sources. This might results in more vague localizations than the comparison algorithm.

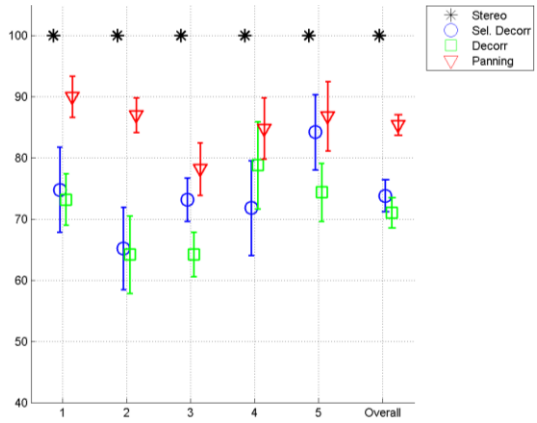

Figure 7. The result of subject test for timbre

For the timbral quality, the re-panning algorithm showed the highest score. As the separated sources are played without any delay or filtering, the artifacts owing to the separation become less audible. Both the comparison and selective decorrelation algorithms suffer from similar amount of timbral degradations to provide similar source widths.

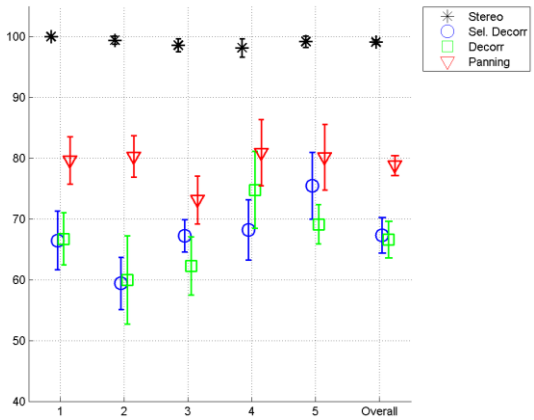

Figure 8. The result of BAQ scores

The BAQ scores show the overall quality scores considering both spatial and timbral qualities. The re-panning approach obtained the highest scores for every items and overall. The preferences between the comparison algorithm and the selective decorrelation approach showed variances among items. In overall score, both of them showed significantly lower scores than the re-panning approach.

5. Conclusions

We have introduced an algorithm for upmixing a mono–to–stereo algorithm for jazz music. We adopted a harmonic and percussive source separation technique, and re-mixing the separated sources on the front side by maximizing the overall source width and equalizing loudness of stereo channels. In the reported experiments, one of the proposed algorithms, re–panning approach, showed advanced quality on both spatial and timbral criteria.

Future work will focus on the generalization of the proposed methods in terms of target signal. As the proposed algorithms performed in off-line, real-time implementation for the algorithm should be realized for practical uses.

Acknowledgements

This research was funded by the MSIP (Ministry of Science, ICT & Future Planning), Korea in the ICT R&D Program 2013.

References

[1] M. R. Schroeder and B. F. Logan, "-Colorless-Artificial Reverberation," *Journal of the Audio Engineering Society*, vol. 9, pp. 192-197, 1961.

[2] D. FitzGerald, "Upmixing from mono-a source separation approach," in *Digital Signal Processing (DSP), 2011 17th International Conference on*, 2011, pp. 1-7.

[3] D. Fitzgerald, "Harmonic/percussive separation using median filtering," 2010.

[4] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Computational Intelligence and Neuroscience*, vol. 2008, 2008.

[5] M. Cobos and J. Lopez, "Resynthesis of wavefield synthesis scenes from stereo mixtures using sound source separation algorithms," *Journal of the Audio Engineering Society*, 2009.

[6] N. Kamado, M. Hirata, H. Saruwatari, and K. Shikano, "Object-based stereo up-mixer for wave field synthesis based on spatial information clustering," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 594-598.

[7] H. Shim, J. S. Abel, and K.-M. Sung, "Stereo Music Source Separation for 3-D Upmixing," in *Audio Engineering Society Convention 127*, 2009.

[8] K. Choi, S. B. Chon, and K. Kang, "Harmonic and Percussive Separation Based on NMF and Tonality Mask," *ETRI Journal*, vol. 34, 2012.

[9] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556-562, 2001.

[10] K. Brandenburg and J. D. Johnston, "Second generation perceptual audio coding: the hybrid code," in *Audio Engineering Society Convention 88*, 1990.

[11] M. J. Hawksford and N. Harris, "Diffuse signal processing and acoustic source characterization for applications in synthetic loudspeaker arrays," in *Audio Engineering Society Convention 112*, 2002.

[12] ANSI, "ANSI S3. 4-2007. Procedure for the computation of loudness of steady sounds," ed: American National Standards Institute New York, 2007.

[13] ITU-R, "Bs. 1534-1,Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," *International Telecommunications Union, Geneva, Switzerland*, 2001.

[14] I.-R. R. BS, "1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," *International Telecommunication Union Abbreviations*, 1997.

[15] M. Cobos and J. J. Lopez, "Resynthesis of Sound Scenes on Wave-Field Synthesis from Stereo Mixtures Using Sound Source Separation Algorithms," *Journal of the Audio Engineering Society*, vol. 57, pp. 91-110, 2009.

[16] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *The Journal of the Acoustical Society of America*, vol. 118, p. 968, 2005.