ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Repeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation

Sunena J. Rajenimbalkar
M.E Student
Dept. of Electronics and Telecommunication,
TPCT'S College of Engineering,
Osmanabad, Maharashtra, India.

Dr. Sudhir S. Kanade, M.E, Ph.D
Head of Department,
Dept. of Electronics and Telecommunication,
TPCT'S College of Engineering,
Osmanabad, Maharashtra, India.

ABSTRACT

Repetition is a core principle in music. This is especially true for popular songs, generally marked by a noticeable repeating musical structure, over which the singer performs varying lyrics. On this basis, we propose a simple method for separating music and voice, by extraction of the repeating musical structure. First, the period of the repeating structure is found. Then, the spectrogram is segmented at period boundaries and the segments are averaged to create a repeating segment model. Finally, each time-frequency bin in a segment is compared to the model, and the mixture is partitioned using binary time-frequency masking by labeling bins similar to the model as the repeating background. This method can improve on the performance of an existing music/voice separation method without requiring particular features or complex frame works.

Index Terms— Music/Voice Separation, Repeating Pattern, Binary Time-Frequency Masking.

1. INTRODUCTION

"Repetition is the basis of music as an art." [1]. A typical piece of popular music has generally an underlying repeating musical structure, with distinguishable patterns periodically repeating at different levels, with possible variations. An important part of music understanding is the identification of those patterns. To visualize repeating patterns, a two- dimensional representation of the musical structure can be calculated by measuring the(dis)similarity between any two instants of the audio. This similarity matrix can be built from the Mel-Frequency Cepstrum Coefficients (MFCC)[4],the spectrogram [8],the chromagram [7],or other features such as the pitch contour (melody) [11] depending on the application, as long as similar sounds yield similarity in the feature space. The similarity matrix can then be used for example to compute a measure of novelty to locate significant changes in the audio [8] or to compute a beat spectrum to characterize the rhythm of the audio [9]. This ability to detect relevant boundaries within the audio can be of great utility for audio segmentation and audio summarization [7],[8],[11].

We propose to apply such a pattern discovery approach for sound separation, by means of extracting the repeating musical structure. The basic idea is to identify in the spectrogram of a song, time-frequency bins that seem to periodically repeat, and extract them using binary time-frequency masking. An immediate application would be music/voice separation.

Music/voice separation systems usually first detect the vocal segments using some features such as MFCCs, and then apply separation techniques such as Non-negative Matrix Factorization [16], pitch-based inference [19],[21], or adaptive Bayesian modeling[18]. Unlike previous approaches, our method does not depend on particular features, does not rely on complex frame works, and does not require prior training. Because it is only based on self-similarity, this method could potentially work on any audio, as long as there is a repeating structure.

It has therefore the advantage of being simple, fast, blind, and also completely automatable.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A. Music Structure Analysis:

In music theory, Schenker asserted that repetition is what gives rise to the concept of the motive, which is defined as the smallest structural element within a musical piece. Ruwetused repetition as a criterion for dividing music into small parts, revealing the syntax of the musical piece. Ockelford argued that repetition/imitation is what brings order to music, and order is what makes music aesthetically pleasing. Bartsch detected choruses in popular music by analyzing the structural redundancy in a similarity matrix built from the chromagram. Other audio thumbnailing methods include Cooper et al. who built a similarity matrix using MFCCs. Dannenberg et al. generated a description of the musical structure related to the AABA form by using similarity matrices built from monophonic pitch estimation, and also the chromagram and a polyphonic transcription. Other music summarization methods include Peeters who built similarity matrices using MFCCs, the chromagram, and dynamic rhythmic features. Foote et al. developed the beat spectrum, a measure of acoustic self-similarity as a function of the time lag, by using a similarity matrix built from the spectrogram. Other beat estimation methods include Pikrakiset al. who built a similarity matrix using MFCCs.

B. Music/Voice Separation:

Music/voice separation methods typically first identify the vocal/non-vocal segments, and then use a variety of techniques to separate the lead vocals from the background accompainment, including spectrogram factorization, accompaniment model learning, and pitch-based inference techniques. Vembu et al. first identified the vocal and non-vocal regions by computing features such as MFCCs, Perceptual Linear Predictive coefficients (PLP), and Log Frequency Power Coefficients (LFPC), and using classifiers such as Neural Networks (NN) and Support Vector Machines (SVM).

They then used Non-negative Matrix Factorization (NMF) to separate the spectrogram into vocal and non-vocal basic components. How-ever, for an effective separation, NMF requires a proper initialization and the right number of components. Raj et al. used a priori known non-vocal segments to train an accompaniment model based on a Probabilistic Latent Component Analysis (PLCA). They then fixed the accompaniment model to learn the vocal parts. Ozerovet al. first performed a vocal/non-vocal segmentation using MFCCs and Gaussian Mixture Models (GMM). They then trained Bayesian models to adapt an accompaniment model learned from the non-vocal segments. However, for an effective separation, such accompaniment model learning techniques require a sufficient amount of non-vocal segments and an accurate vocal/ non-vocal prior segmentation. Hsu et al. first used a Hidden Markov Model (HMM) to identify accompaniment, voiced, and unvoiced segments.

They then used the inference method of Li et al. to separate the voiced vocals, while the pitch contour was derived from the predominant pitch estimation algorithm of Dressler. In addition, they proposed a method to separate the unvoiced vocals based on GMMs and a method to enhance the voiced vocals based on spectral subtraction.

The rest of the paper is organized as follows. Section2 presents the method. Result analysis is done in Section3. Finally, conclusion and perspectives are discussed in Section4.

2. PROPOSED METHOD

Repeating Pattern Extraction Technique(REPET):

Repetition in each music structure is its basic principle. Any musical pieces being characterized by an underlying repetitive structure over which varying elements are superimposed.

The basic idea is to:
A. Identify the periodically repeating segments,
B. Repeating segment modeling, and
C. Extract the repeating patterns via time-frequency masking.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

2.1. Identify the periodically repeating segments:

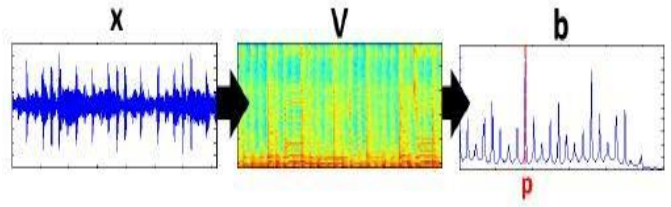


Fig.1. Calculation of the beat spectrum and estimation of the repeating period. Repeating period p from the beat spectrum b.

Periodicities in any mixture signal can be found by using the autocorrelation, measuring the similarities between segments and lagged version of itself over the successive intervals of time.

Given a mixture signal x, method first calculate its Short- Time Fourier Transform (STFT) X, by using half-overlapping Hamming windows of N samples. Then derives magnitude spectrogram V by taking absolute values of the elements of X, while keeping the DC component and discarding the symmetric part of segment. Then computing autocorrelation of each row of the power spectrogram V^2 (element-wise square of V) and obtain the matrix B. Then useV^2 to emphasize the appearance of peaks of periodicity in B. If the mixture signal x is stereo, then averaging of V^2 over the channels. The overall acoustic self-similarity b of x is obtained by taking the mean over the rows of B. Then finally normalizes b by its first term (lag0).

2.2. Repeating Segment Model:

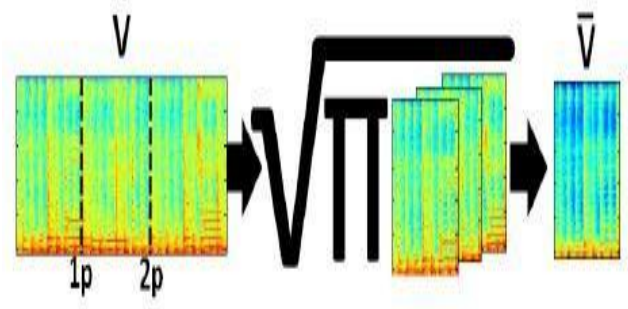


Fig.2. Segmentation of the mixture spectrogram and computation of the repeating segment model. Segmentation of V to get the mean repeating segment $\bar{V}$.

After estimation the period p of the repeating musical structure, the method uses it to evenly segment the spectrogram V into segments of length p. Then computing mean repeating segment $\bar{V}$ over r portion of V, which can be thought of as the repeating segment model. The approach is that time-frequency bins comprises the repeating patterns had similar values at each period that would also be similar to the repeating segment model. Experiments had shown that the geometric mean lead to a effective extraction of the repeating musical structure than arithmetic mean.

2.3.Binary Time-Frequency Masking:

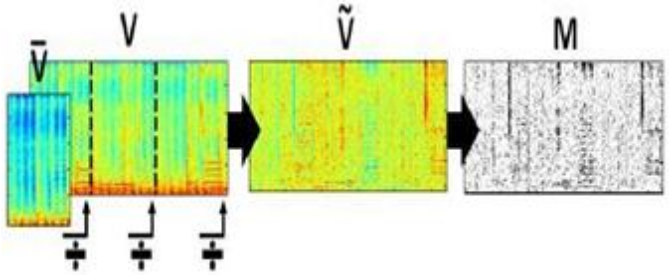


Fig.3. Derivation of the repeating spectrogram model and building of the soft time-frequency mask. Bin-wise division of V by $\bar{V}$to get the binary time-frequency mask M.

After computing the mean repeating segment $\bar{V}$, method divides each time-frequency bin in each segment of spectrogram V by the corresponding bin in $\bar{V}$. Then taking the absolute value of the logarithm of each bin to get a modified spectrogram $\tilde{V}$ and furthermore the repeating musical structure generally involving variations. Therefore, method introduce a tolerance t when creating the binary time frequency mask M. Experiments shows tolerance of t =1 giving good separation results, both for music and voice.

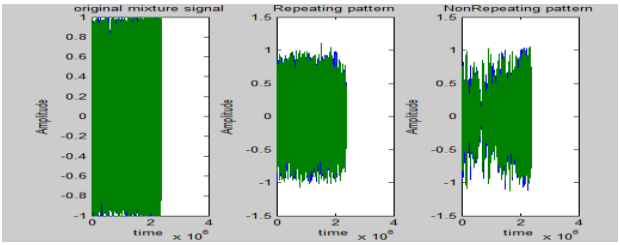
Once the binary time-frequency mask M is computed, then symmetrising and applying to STFT X of the mixture signal x to have the STFT of the music and the STFT of the speech. The music signal and voice are finally achieved by inverting their corresponding STFTs into the time domain.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

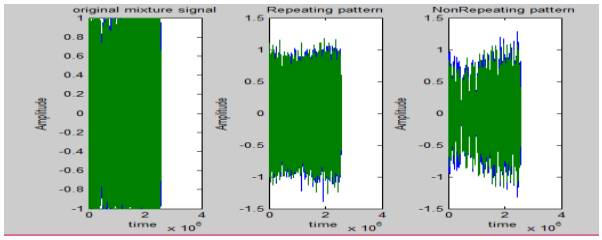
3. RESULTS

We evaluated our music/voice separation system using various song clips at a sample rate of 44.1 kHz, with duration ranging from 20 to 30 sec. In the separation process, the STFT of each mixture x was calculated using a half-overlapping Hamming window. The repeating period p was automatically estimated from the beat spectrum b simply by computing the local maxima in b and identifying the one that periodically repeats the most often, with the highest accumulated energy over its periods. When building the binary time-frequency mask, we fixed the tolerance t to 1. Our music/ voice separation system is thus completely automatic.

Fig.4. shows waveform comparison of original mixture audio, repeating audio, and non-repeating audio. The vertical axis in each plot indicates the amplitude of the waveform and horizontal axis indicates time.

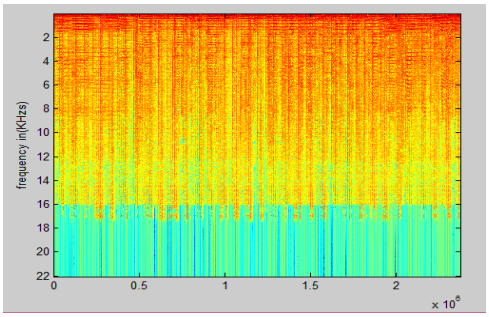


Result for song clip 1

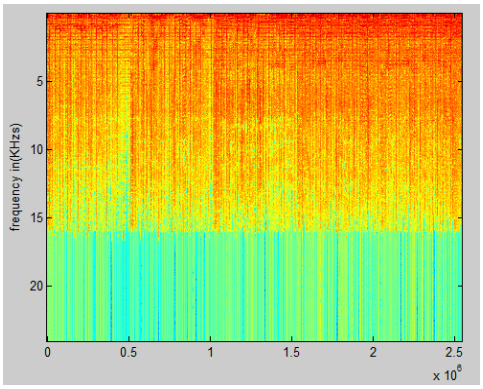


Result for song clip 2

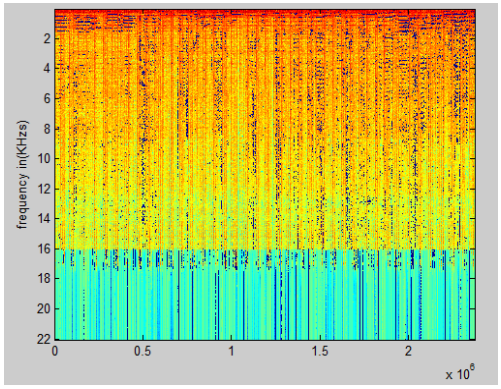
Fig.4. Waveform comparison**:** original mixture audio, repeating audio and non-repeating audio.



(a)Original mixture audio spectrogram

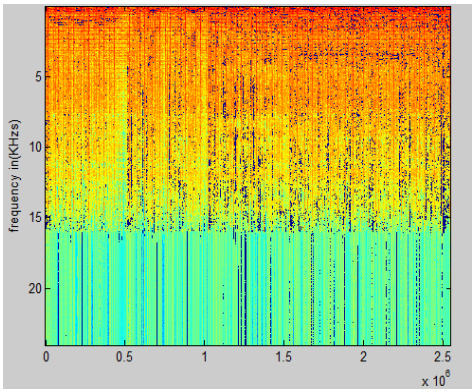


(a)Original mixture audio spectrogram

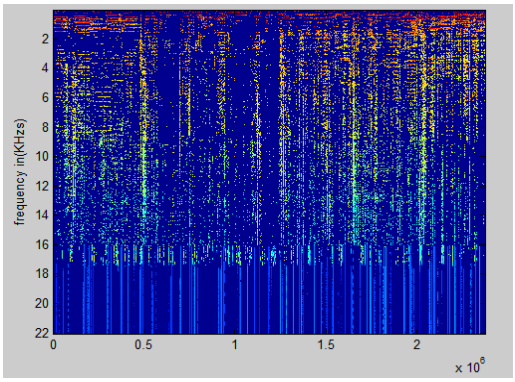


(b) Repeating audio spectrogram

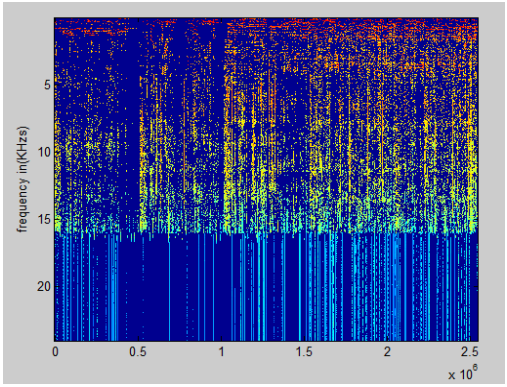
ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal



(b) Repeating audio spectrogram



(c) Non-repeating audio spectrogram



(c) Non-repeating audio spectrogram

Result for song clip 1 Result for song clip 2

Fig.5. Spectrogram comparison**:** (a) Original mixture audio (b) Repeating audio (c) Non-repeating audio.

Fig. 5(a) shows the spectrogram of original mixture audio and fig.5(b) and fig.5(c) shows the spectrogram estimated by the separation system i.e. repeating audio and non-repeating audio. The vertical axis in each plot indicates the frequency of the waveform and horizontal axis indicates time. As shown in fig.5 the non-repeating foreground (voice) has a sparse and varied time-frequency representation compared with the time frequency representation of the repeating background (music) – a reasonable assumption for voice in music, time-frequency bins with little deviation at period p would constitute a repeating pattern.

4. CONCLUSION

We have proposed a novel method for music/voice separation, by extraction of the underlying musical repeating structure. This method can achieve better separation performance than an existing automatic approach, without requiring particular features or complex frameworks. This method also has the advantage of being simple, fast and completely automatable.

There are several directions in which we want to take this work. First, we would like to improve our automatic music/voice separation system by (1) implementing a better repeating period finder, (2) building better time-frequency masks, for example by using a measure of repetitiveness when assigning time-frequency bins, and (3) taking into account the pitch, timbre, or multichannel information. We could also combine our method with other existing music/voice separation systems to improve separation performance. Then, we would like to extend this separation approach for the extraction of multiple hierarchical repeating structures, by using repeating periods at different levels. Finally, we would like to apply this separation approach to the extraction of individual repeating patterns by using a similarity matrix. This could be used for the separation of structural elements in music.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

REFERENCES

[1] H. Schenker, Harmony. Chicago, IL: Univ. of Chicago Press, 1954.

[2] N. Ruwet and M. Everist, "Methods of analysis in musicology," MusicAnal., vol. 6, no. 1/2, pp. 3–9+11–36, Mar.-Jul. 1987.

[3] A. Ockelford, Repetition in Music: Theoretical and MetatheoreticalPerspectives. Farnham, U.K.: Ashgate, 2005, vol. 13, Royal MusicalAssociation Monographs.

[4] J. Foote, "Visualizing music and audio using self-similarity," in Proc. 7th ACM nt. Conf. Multimedia (Part 1), Orlando, FL, Oct.-Nov. 30–05,1999, pp. 77–80.

[5] M. Cooper and J. Foote, "Automatic music summarization via similarityanalysis," in Proc. 3rd Int. Conf. Music Inf. Retrieval, Paris,France, Oct. 13–17, 2002, pp. 81–85.

[6] A. Pikrakis,I. Antonopoulos, and S. Theodoridis, "Music meter andtempo tracking from raw polyphonic audio," in Proc. 9th Int. Conf.Music Inf. Retrieval, Barcelona, Spain, Oct. 10–14, 2008.

[7] G. Peeters, "Deriving musical structures from signal analysis for musicaudio summary generation: "Sequence" and "state" approach," in ComputerMusic Modeling and Retrieval, U.Wiil, Ed. Berlin/Heidelberg,Germany: Springer, 2004, vol. 2771, Lecture Notes in Computer Science,pp. 169–185.

[8] J. Foote, "Automatic audio segmentation using a measure of audionovelty," in Proc. IEEE Int. Conf. Multimedia and Expo, New York,Jul.-Aug. 30–02, 2000, vol. 1, pp. 452–455.

[9] J. Foote and S. Uchihashi, "The beat spectrum: A new approach torhythm analysis," in Proc. IEEE Int. Conf. Multimedia and Expo,Tokyo, Japan, Aug. 22–25, 2001, pp. 881–884.

[10] M. A. Bartsch,"To catch a chorus using chroma-based representationsfor audio thumbnailing," in Proc. IEEE Workshop Applicat. SignalProcess. Audio Acoust., New Paltz, NY, Oct. 21–24, 2001, pp. 15–18.

[11] R. B. Dannenberg and N. Hu, "Pattern discovery techniques for musicaudio," J. New Music Res., vol. 32, no. 2, pp. 153–164, 2003.

[12] K. Jensen, "Multiple scale music segmentation using rhythm, timbre,and harmony," EURASIP J. Adv. Signal Process., vol. 2007, no. 1, pp.1–11, Jan. 2010.

[13] R. B. Dannenberg, "Listening to "Naima": An automated structuralanalysis of music from recorded audio," in Proc. Int. Comput. MusicConf., Gothenburg, Sweden, Sep. 17–21, 2002, pp. 28–34.

[14] R. B. Dannenberg and M. Goto, "Music structure analysis fromacoustic signals," in Handbook of Signal Processing in Acoustics, D.Havelock, S. Kuwano, and M. Vorländer, Eds. New York: Springer,2009, pp. 305–331.

[15] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structureanalysis," in Proc. 11th Int. Soc. Music Inf. Retrieval, Utrecht, TheNetherlands, Aug. 9–13, 2010, pp. 625–636.

[16] S. Vembu and S. Baumann, "Separation of vocals from polyphonicaudio recordings," in Proc. 6th Int. Conf. Music Inf. Retrieval, London,U.K., Sep. 11–15, 2005, pp. 337–344.

[17] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foregroundsinger from background music," in Proc. Int. Symp. Frontiersof Res. Speech and Music, Mysore, India, May 8–9, 2007.

[18] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation ofBayesian models for single-channel source separation and its applicationto voice/music separation in popular songs," IEEE Trans.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Audio,Speech, Lang. Process., vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[19] Y. Li and D. Wang, "Separation of singing voice from music accompanimentfor monaural recordings," IEEE Trans. Audio, Speech, Lang.Process., vol. 15, no. 4, pp. 1475–1487, May 2007.

[20] M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri, "Accompanimentseparation and karaoke application based on automatic melodytranscription," in Proc. IEEE Int. Conf. Multimedia & Expo, Hannover,Germany, Jun. 23–26, 2008, pp. 1417–1420.

[21] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-basedinference and non-negative spectrogram factorization in separating vocalsfrom polyphonic music," in ISCA Tutorial and ResearchWorkshopon Statistical and Perceptual Audition, Brisbane, Australia, Sep. 21,2008, pp. 17–20.

[22] K. Dressler,"An auditory streaming approach on melody extraction,"in Proc. 7th Int. Conf. Music Inf. Retrieval (MIREX Eval.), Victoria,BC, Canada, Oct. 8–12, 2006.

[23] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separationfor monaural recordings using the MIR-1K dataset," IEEE Trans.Audio, Speech, Lang. Process., vol. 18, no. 2, pp. 310–319, Feb. 2010.

[24] J.-L. Durrieu, B. David, and G. Richard, "A musically motivatedmid-level representation for pitch estimation and musical audio sourceseparation," IEEE J. Sel. Topics Signal Process., vol. 5, no. 6, pp.1180–1191, Oct. 2011.

[25] M. Piccardi, "Background subtraction techniques: A review," in Proc.IEEE Int. Conf. Syst., Man, Cybern., The Hague, The Netherlands, Oct.10–13, 2004, pp. 3099–3104.

[26] ZafarRafii and Bryan Pardo, "REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation", IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 1, pp. 71 – 82, January 2013.

[27] Z. Rafii and B. Pardo, "A simple music/voice separation system based on the extraction of the repeating musical structure," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Prague, Czech Republic, May 22–27, 2011, pp. 221–224.

[28] K. Yoshii, M. Goto, and H. G. Okuno, "Adamast: A drum sound recognizer based on adaptation and matching of spectrogram templates," in Proc. 5th Int. Conf. Music Inf. Retrieval, Barcelona, Spain, Oct. 10–14, 2004, pp. 184–191.

[29] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hean, J. R. Zeidler, J. E. Dong, and R. C. Goodlin, "Adaptive noise cancelling: Principles and applications," Proc. IEEE, vol. 63, no. 12, pp. 1692–1716, Dec. 1975.

[30] J. H. McDermott, D. Wrobleski, and A. J. Oxenham, "Recovering sound sources from embedded repetition," Proc. Natural Acad. Sci.United States of Amer., vol. 108, no. 3, pp. 1188–1193, Jan. 2011.