ZE AWESOME FRENCH AUDIO RESEARCHER (ZAFAR)

Zafar Rafii

PhD in Electrical Engineering & Computer Science
zafarrafii@gmail.com

ABSTRACT

We present Zafar, Ze Awesome French Audio Researcher. The proposed researcher has a PhD in electrical engineering and computer science from Northwestern University, with a focus on audio signal analysis. He has over 30 publications, including conference papers, journal articles, and patents, with 1,500 citations overall. He is actively involved within the research community, as a reviewer for numerous conferences and journals, a member of the IEEE audio and acoustic signal processing technical committee, and an organizer of networking meetups in the San Francisco Bay Area. He is currently a research engineer manager at Gracenote, where he is working on a number of projects involving, among others, audio recognition, audio separation, and audio classification.

Index Terms— Research, audio signal analysis, separation, recognition, classification

1. INTRODUCTION

The proposed researcher is named Zafar Rafii. He received a PhD in electrical engineering and computer science from Northwestern University in 2014. He was with the Interactive Audio Lab under the supervision of professor Bryan Pardo. Prior to that, he was a research engineer at Audionamix, in France. He is now a research engineer manager in the Applied Research group at Gracenote.

Fig. 1. Overview of the proposed researcher.

The proposed researcher has interest and expertise in audio signal analysis. He has worked on a number of projects, including:

- Blind source separation
- Spatial source separation
- Digital audio effects
- Audio fingerprinting
- Cover song identification
- Audio encoding analysis
- Audio beamforming
- Audio watermarking
- Audio/video segmentation
- Audio classification

For more information on the proposed researcher, the reader is referred to the following materials:

- CV
- GitHub
- LinkedIn
- Google Scholar

For other relevant information related to the proposed researcher, such as the meetups he organizes, the mentoring program he is involved in, or the audio dataset he created, the reader is referred to the following links:

- SF-BISH Bash meetup
- SF-BISH Bash YouTube channel
- Women in Music Information Retrieval
- MUSDB18 dataset

The rest of the website is organized as follows. In Section 2, we present a selection of projects the proposed researcher has worked on. In Section 3, we introduce his PhD thesis work on the REpeating Pattern Extraction Technique (REPET) for blind source separation. In Section 4, we share links to his GitHub repositories where some of his source codes reside. In Section 5, we provide references to all of his publications, presentations, and other materials.

2. RESEARCH

2.1. Adaptive Reverberation Tool (2008)

People often think about sound in terms of subjective concepts which do not necessarily have known mappings onto the controls of existing audio tools. For example, a bass player may wish to use a reverberation effect to make her/his bass sound more ”boomy”, but unfortunately there is no ”boomy”

knob to be found. We developed a system that can quickly learn an audio concept from a user (e.g., a "boomy" effect) and generate a simple controller than can manipulate sounds in terms of that audio concept (e.g., make a sound more "boomy"), bypassing the bottleneck of technical knowledge of complex interfaces and individual differences in subjective terms.

For this study, we focused on reverberation effects. We developed a digital reverberator, mapping the parameters of the digital filters to measures of the reverberation effect, so that the reverberator can be controlled through meaningful descriptors such as "reverberation time" or "spectral centroid." In the learning process, a sound is first modified by a series of reverberation settings using the reverberator. The user then listens and rates each modified sound as to how well it fits the audio concept she/he has in mind. The ratings are finally mapped onto the controls of the reverberator and a simple controller is built with which the user is able to manipulate the degree of her/his audio concept on a sound. Several experiments conducted on human subjects showed that the system learns quickly (under 3 minutes), predicts user responses well (mean correlation of 0.75), and meets users' expectations (average human rating of 7.4 out of 10).

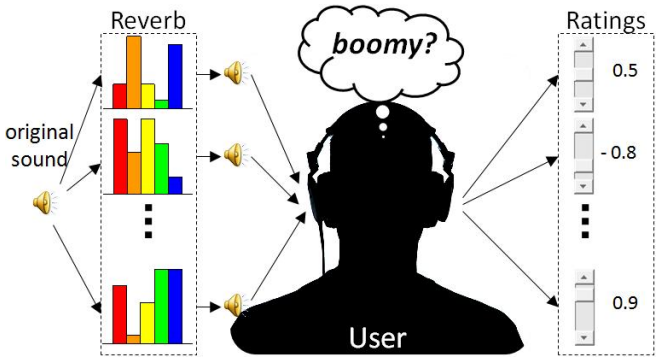

Fig. 2. A listener rating a sound modified by a series of reverberation settings as to how well it fits the audio concept of "boomy" they have in mind.

For more information about this project, the reader is referred to [17], [36], and [39].

2.2. DUET using the CQT (2011)

The Degenerate Unmixing Estimation Technique (DUET) is a blind source separation method that can separate an arbitrary number of unknown sources using a single stereo mixture. DUET builds a two-dimensional histogram from the amplitude ratio and phase difference between channels, where each peak indicates a source, with peak location corresponding to the mixing parameters associated with that source. Provided that the time-frequency bins of the sources do not overlap too much - an assumption generally validated by speech mixtures, DUET partitions the time-frequency representation of the mixture by assigning each bin to the source with the closest mixing parameters. However, when time-frequency bins of the sources start to overlap more - as generally seen in music mixtures when using the common short-time Fourier transform (STFT), peaks start to fuse in the 2d histogram and DUET cannot perform separation effectively.

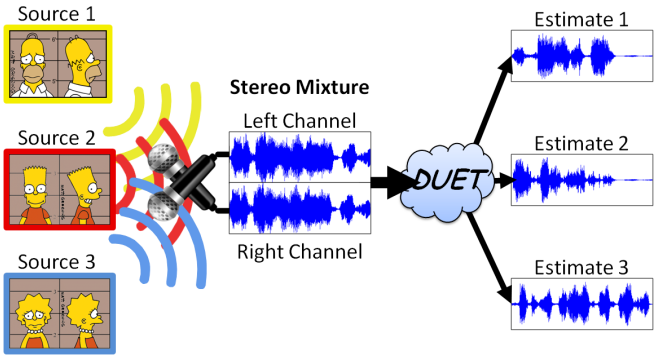

Fig. 3. Blind source separation of a stereo recording of Homer, Bart, and Lisa using DUET.

We proposed to improve peak/source separation in DUET by building the 2d histogram from an alternative time-frequency representation based on the constant-Q transform (CQT). Unlike the Fourier transform, the CQT has a logarithmic frequency resolution, mirroring the human auditory system and matching the geometrically spaced frequencies of the Western music scale, therefore better adapted to music mixtures. We also proposed other contributions to enhance DUET, such as adaptive boundaries for the 2d histogram to improve peak resolving when sources are spatially too close to each other, and Wiener filtering to improve source reconstruction. Experiments on mixtures of piano notes and harmonic sources showed that peak/source separation is overall improved, especially at low octaves (under 200 Hz) and for small mixing angles (under $\pi/6$ rad).

Unlike the classic DUET based on the Fourier transform, DUET combined with the CQT can resolve adjacent pitches in low octaves as well as in high octaves thanks to the log frequency resolution of the CQT:

- Mixture of 3 piano notes
- Estimates: A2 - Bb2 - B2
- Originals: A2 - Bb2 - B2

DUET combined with the CQT and adaptive boundaries helps to improve separation when sources have low pitches (for example, here, between the two cellos) and/or are spatially too close to each other:

- Mixture of 4 instruments
- Estimates: cello 1 - cello 2 - flute - strings
- Originals: cello 1 - cello 2 - flute - strings

For more information about this project, the reader is referred to [35].

2.3. Live Music Fingerprinting (2014)

Suppose that you are at a music festival checking on an artist, and you would like to quickly know about the song that is being played (e.g., title, lyrics, album, etc.). If you have a smartphone, you could record a sample of the live performance and compare it against a database of existing recordings from the artist. Services such as Shazam or SoundHound will not work here, as this is not the typical framework for audio fingerprinting or query-by-humming systems, as a live performance is neither identical to its studio version (e.g., variations in instrumentation, key, tempo, etc.) nor it is a hummed or sung melody. We propose an audio fingerprinting system that can deal with live version identification by using image processing techniques. Compact fingerprints are derived using a log-frequency spectrogram and an adaptive thresholding method, and template matching is performed using the Hamming similarity and the Hough Transform.

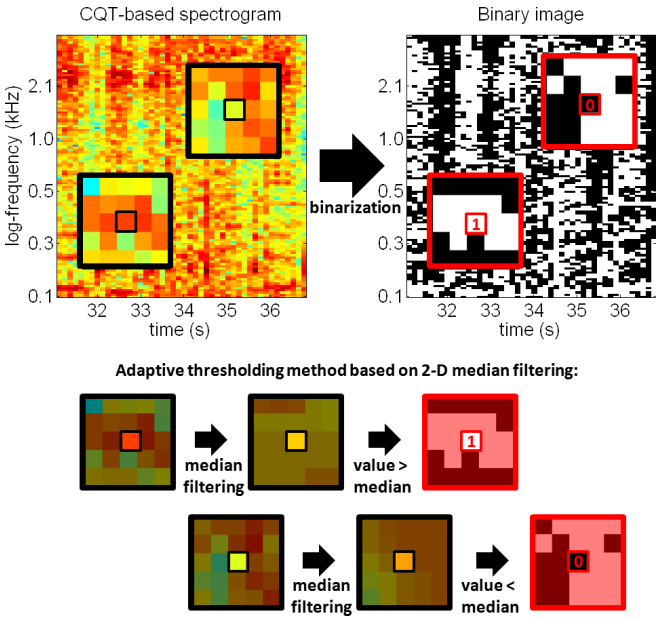

Fig. 4. Overview of the fingerprinting stage. The audio signal is first transformed into a log-frequency spectrogram by using the CQT. The CQT-based spectrogram is then transformed into a binary image by using an adaptive thresholding method.

For more information about this project, the reader is referred to [28].

2.4. Lossy Audio Compression Identification (2018)

We propose a system which can estimate from an audio recording that has previously undergone lossy compression the parameters used for the encoding, and therefore identify the corresponding lossy coding format. The system analyzes the audio signal and searches for the compression parameters and framing conditions which match those used for the encoding. In particular, we propose a new metric for measuring traces of compression which is robust to variations in the audio content and a new method for combining the estimates from multiple audio blocks which can refine the results. We evaluated this system with audio excerpts from songs and movies, compressed into various coding formats, using different bit rates, and captured digitally as well as through analog transfer. Results showed that our system can identify the correct format in almost all cases, even at high bitrates and with distorted audio, with an overall accuracy of 0.96.

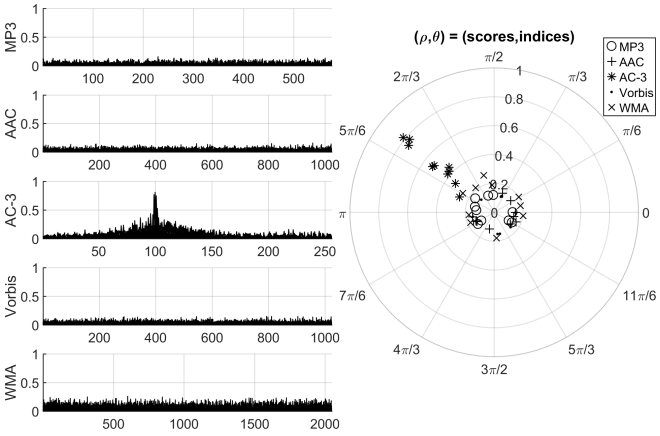

Fig. 5. Results for an audio example encoded with AC-3. The system identified traces of compression corresponding to AC-3, but not to other lossy coding formats such as MP3, AAC, Vorbis, or WMA.

For more information about this project, the reader is referred to [19].

2.5. Sliding DFT with Kernel Windowing (2018)

The sliding discrete Fourier transform (SDFT) is an efficient method for computing the N-point DFT of a given signal starting at a given sample from the N-point DFT of the same signal starting at the previous sample. However, the SDFT does not allow the use of a window function, generally incorporated in the computation of the DFT to reduce spectral leakage, as it would break its sliding property. We show how windowing can be included in the SDFT by using a kernel derived from the window function, while keeping the process computationally efficient. In addition, this approach allows for turning other transforms, such as the modified discrete cosine transform (MDCT), into efficient sliding versions of themselves.

For more information about this project, the reader is referred to [12].

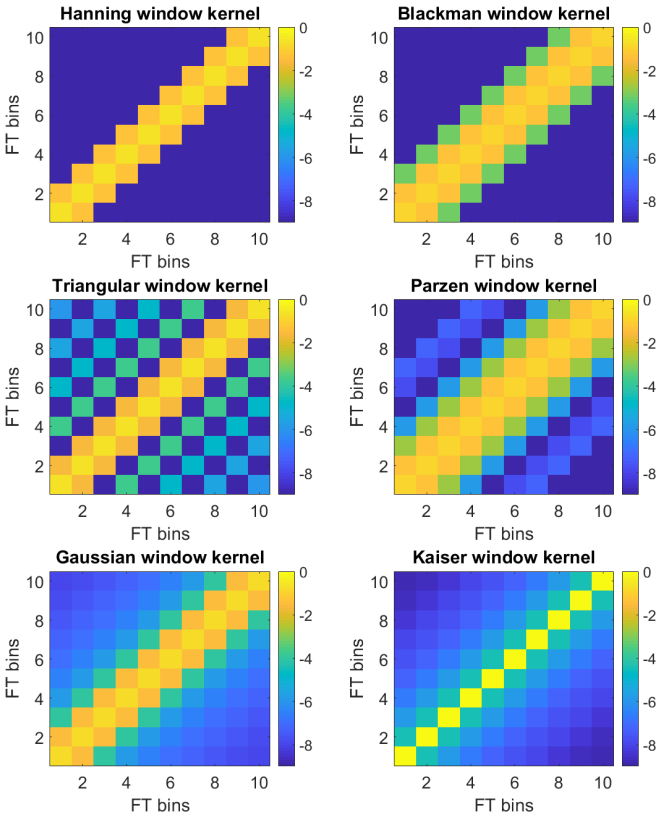

Fig. 6. Kernels derived from the (a) Hanning, (b) Blackman, (c) triangular, (d) Parzen, (e) Gaussian (with $\alpha = 2.5$), and (f) Kaiser (with $\beta = 0.5$) windows. The kernels were derived for an N-point DFT where N = 2,048 samples. Only the first 100 coefficients at the bottom-left corner of the N-by-N kernels are shown. The values are displayed in log of amplitude.

3. REPET

Repetition is a fundamental element in generating and perceiving structure. In audio, mixtures are often composed of structures where a repeating background signal is superimposed with a varying foreground signal. On this basis, we present the REpeating Pattern Extraction Technique (REPET), a simple approach for separating the repeating background from the non-repeating foreground in an audio mixture. The basic idea is to find the repeating elements in the mixture, derive the underlying repeating models, and extract the repeating background by comparing the models to the mixture. Unlike other separation approaches, REPET does not depend on special parameterizations, does not rely on complex frameworks, and does not require external information. Because it is only based on repetition, it has the advantage of being simple, fast, blind, and therefore completely and easily automatable.

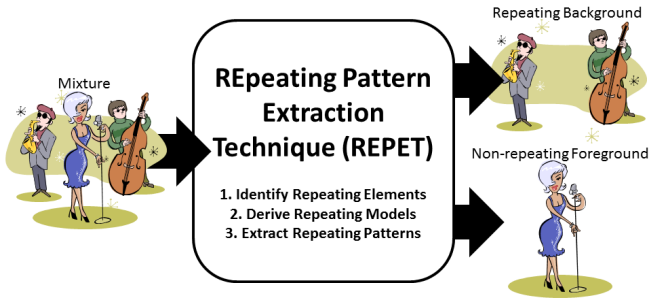

Fig. 7. Overview of REPET.

3.1. Original REPET (2011)

The original REPET aims at identifying and extracting the repeating patterns in an audio mixture, by estimating a period of the underlying repeating structure and modeling a segment of the periodically repeating background.

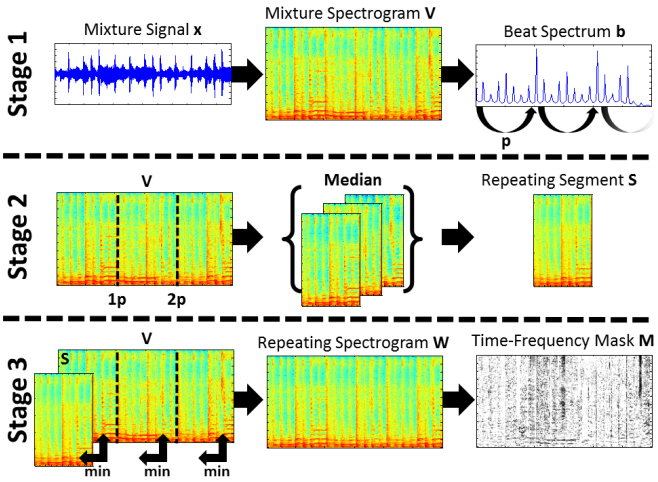

Fig. 8. Overview of the original REPET. Stage 1: calculation of the beat spectrum b and estimation of a repeating period p. Stage 2: segmentation of the mixture spectrogram V and calculation of the repeating segment S. Stage 3: calculation of the repeating spectrogram W and derivation of the time-frequency mask M.

Experiments on a data set of song clips showed that REPET can be effectively applied for music/voice separation. Experiments also showed that REPET can be combined with other methods to improve background/foreground separation; for example, it can be used as a preprocessor to pitch detection algorithms to improve melody extraction, or as a postprocessor to a singing voice separation algorithm to improve music/voice separation.

- Mixture
- Estimates: background - foreground
- Originals: accompaniment - vocals

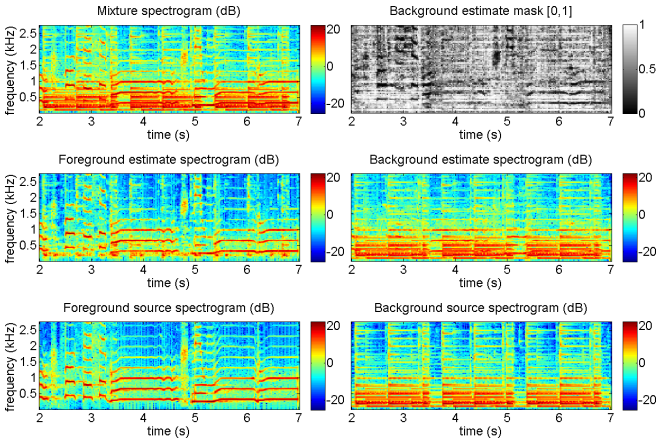

Fig. 9. Music/voice separation using REPET. The mixture is a female singer (foreground) singing over a guitar accompaniment (background). The guitar has a repeating chord progression that is stable along the song. The spectrograms and the mask are shown for 5 seconds and up to 2.5 kHz.

REPET can be easily extended to handle varying repeating structures, by simply applying the method along time, on individual segments or via a sliding window. Experiments on a data set of full-track real-world songs showed that this method can be effectively applied for music/voice separation.

For more information about this project, the reader is referred to [34], [16], and [38].

3.2. Adaptive REPET (2012)

The original REPET works well when the repeating background is relatively stable (e.g., a verse or the chorus in a song); however, the repeating background can also vary over time (e.g., a verse followed by the chorus in the song). The adaptive REPET is an extension of the original REPET that can handle varying repeating structures, by estimating the time-varying repeating periods and extracting the repeating background locally, without the need for segmentation or windowing.

Experiments on a data set of full-track real-world songs showed that the adaptive REPET can be effectively applied for music/voice separation.

- Mixture
- Estimates: background - foreground
- Originals: accompaniment - vocals

For more information about this project, the reader is referred to [32] and [38].

3.3. REPET-SIM (2012)

The REPET methods work well when the repeating background has periodically repeating patterns (e.g., jackhammer noise); however, the repeating patterns can also happen intermittently or without a global or local periodicity (e.g., frogs by a pond). REPET-SIM is a generalization of REPET that can also handle non-periodically repeating structures, by using a similarity matrix to identify the repeating elements.

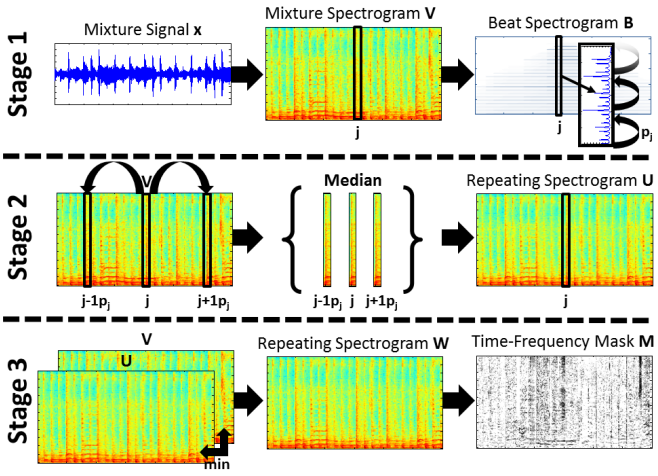

Fig. 10. Overview of the adaptive REPET. Stage 1: calculation of the beat spectrogram B and estimation of the repeating periods pj's. Stage 2: filtering of the mixture spectrogram V and calculation of an initial repeating spectrogram U. Stage 3: calculation of the refined repeating spectrogram W and derivation of the time-frequency mask M.

Experiments on a data set of full-track real-world songs showed that REPET-SIM can be effectively applied for music/voice separation.

REPET-SIM can be easily implemented online to handle real-time computing, particularly for real-time speech enhancement. The online REPET-SIM simply processes the time frames of the mixture one after the other given a buffer that temporally stores past frames. Experiments on a data set of two-channel mixtures of one speech source and real-world background noise showed that the online REPET-SIM can be effectively applied for real-time speech enhancement.

- Mixture
- Estimates: foreground - background
- Originals: speech - noise

For more information about this project, the reader is referred to [31], [30], and [38].

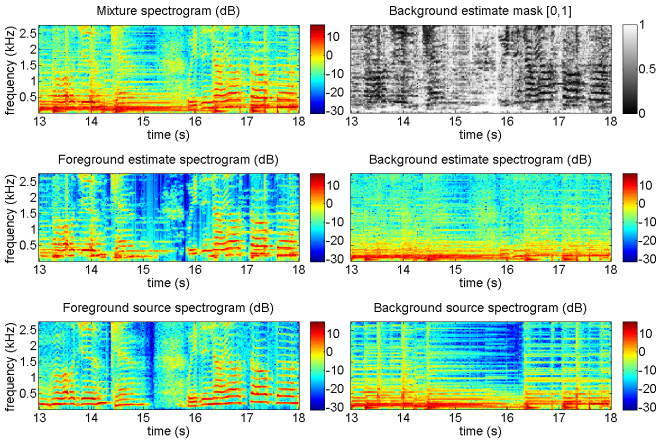

Fig. 11. Music/voice separation using the adaptive REPET. The mixture is a male singer (foreground) singing over a guitar and drums accompaniment (background). The guitar has a repeating chord progression that changes around 15 seconds. The spectrograms and the mask are shown for 5 seconds and up to 2.5 kHz.

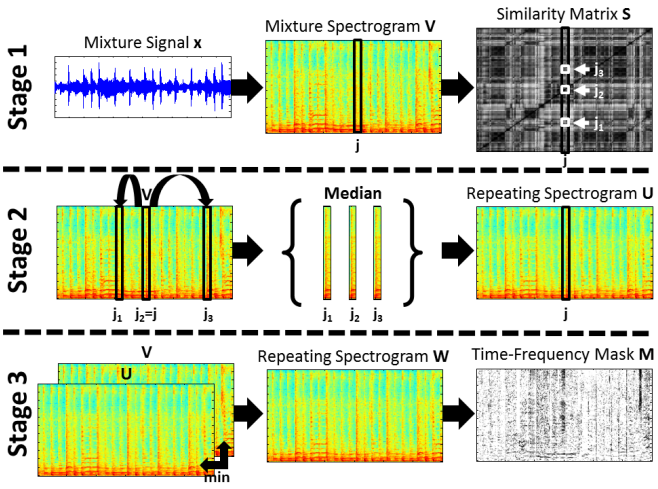

Fig. 12. Overview of REPET-SIM. Stage 1: calculation of the similarity matrix S and estimation of the repeating indices jk's. Stage 2: filtering of the mixture spectrogram V and calculation of an initial repeating spectrogram U. Stage 3: calculation of the refined repeating spectrogram W and derivation of the time-frequency mask M.

3.4. uREPET (2015)

3.5. PROJET-MAG (2017)

4. CODES

5. REFERENCES

5.1. Patents

[1] M. Cremer, R. Coover, Z. Rafii, A. Vartakavi, A. Schmidt, and T. Hodges, "Methods and Apparatus to Control Lighting Effects," 16/698,697, May 2021.

[2] Z. Rafii, "Methods and Apparatus to Extract a Pitch-independent Timbre Attribute from a Media Signal," 16/698,697, May 2021.

[3] M. Cremer, Z. Rafii, R. Coover, and P. Seetharaman, "Automated Cover Song Identification," 17/065,479, Feb. 2021.

[4] Z. Rafii, M. Cremer, and B. Kim, "Methods, Apparatus and Articles of Manufacture to Identify Sources of Network Streaming Services," 16/984,091, Jan. 2021.

[5] Z. Rafii, "Methods and Apparatus to Improve Detection of Audio Signatures," 16/455,025, Dec. 2020.

[6] Z. Rafii and P. Seetharaman, "Audio Identification based on Data Structure," 16/927,577, Oct. 2020.

[7] Z. Rafii, "Audio Matching based on Harmonogram," 16/913,162, Oct. 2020.

[8] ——, "Methods and Apparatus to Perform Windowed Sliding Transforms," 15/899,220, Jul. 2020.

[9] Z. Rafii, M. Cremer, and B. Kim, "Methods and Apparatus to Identify Sources of Network Streaming Services using Windowed Sliding Transforms," 16/843,582, Jul. 2020.

[10] R. Coover and Z. Rafii, "Methods and Apparatus to Fingerprint an Audio Signal via Normalization," 16/453,654, Mar. 2020.

[11] B. Pardo and Z. Rafii, "Audio Separation System and Method," 13/612,413, Jul. 2015.

5.2. Journal Articles

[12] Z. Rafii, "Sliding Discrete Fourier Transform with Kernel Windowing," *IEEE Signal Processing Magazine*, vol. 35, no. 6, Nov. 2018.

[13] Z. Rafii, A. Liutkus, F.-R. Stöter, D. F. Stylianos Ioannis Mimilakis, and B. Pardo, "An Overview of Lead and Accompaniment Separation in Music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 26, Aug. 2018.

[14] Z. Rafii, Z. Duan, and B. Pardo, "Combining Rhythm-based and Pitch-based Methods for Background and Melody Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, Dec. 2014.

[15] A. Liutkus, D. FitzGerald, Z. Rafii, B. Pardo, and L. Daudet, "Kernel Additive Models for Source Separation," *IEEE Transactions on Signal Processing*, vol. 62, no. 16, Aug. 2014.

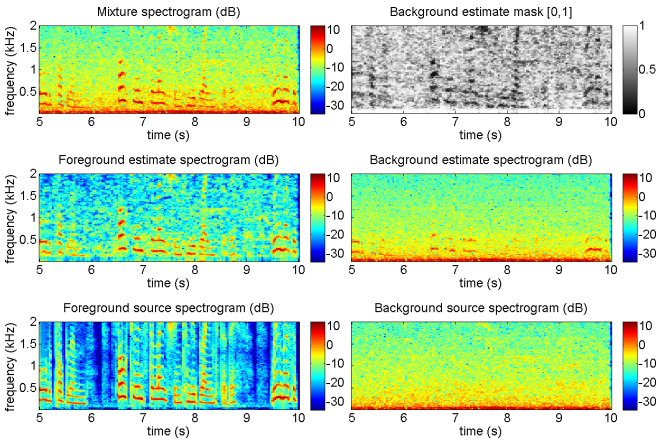

Fig. 13. Noise/speech separation using REPET-SIM. The mixture is a female speaker (foreground) speaking in a town square (background). The square has repeating noisy elements (passers-by and cars) that happen intermittently. The spectrograms and the mask are shown for 5 seconds and up to 2 kHz.

[16] Z. Rafii and B. 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, Jan. 2013.

[17] A. T. Sabin, Z. Rafii, and B. Pardo, "Weighting-Function-Based Rapid Mapping of Descriptors to Audio Processing Parameters," *Journal of the Audio Engineering Society*, vol. 59, no. 6, Jun. 2011.

5.3. Conference Proceedings

[18] A. Vartakavi, A. Garg, and Z. Rafii, "Audio Summarization for Podcasts," in *29th European Signal Processing Conference*, Dublin, Ireland, Aug. 2021, (poster).

[19] B. Kim and Z. Rafii, "Lossy Audio Compression Identification," in *26th European Signal Processing Conference*, Rome, Italy, Sep. 2018, (poster).

[20] P. Seetharaman and Z. Rafii, "Cover Song Identification with 2d Fourier Transform Sequences," in *42nd IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, Mar. 2017, (poster).

[21] D. FitzGerald, Z. Rafii, and A. Liutkus, "User Assisted Separation of Repeating Patterns in Time and Frequency using Magnitude Projections," in *42nd IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, USA, Mar. 2017, (poster).

[22] A. Liutkus, F.-R. Stöter, Z. Rafii, *et al.*, "The 2016 Signal Separation Evaluation Campaign," in *13th International Conference on Latent Variable Analysis and Signal Separation*, Grenoble, France, Feb. 2017.

[23] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 Signal Separation Evaluation Campaign," in *12th International Conference on Latent Variable Analysis and Signal Separation*, Liberec, Czech Republic, Aug. 2015.

[24] Z. Rafii, A. Liutkus, and B. Pardo, "A Simple User Interface System for Recovering Patterns Repeating in Time and Frequency in Mixtures of Sounds," in *40th IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015, (poster).

[25] A. Liutkus, D. FitzGerald, and Z. Rafii, "Scalable Audio Separation with Light Kernel Additive Modelling," in *40th IEEE International Conference on Acoustics, Speech and Signal Processing*, Brisbane, QLD, Australia, Apr. 2015, (slides).

[26] D. FitzGerald, A. Liutkus, Z. Rafii, B. Pardo, and L. Daudet, "Harmonic/Percussive Separation using Kernel Additive Modelling," in *25th IET Irish Signals and Systems Conference*, Limerick, Ireland, Jun. 2014.

[27] A. Liutkus, Z. Rafii, B. Pardo, D. FitzGerald, and L. Daudet, "Kernel Spectrogram Models for Source Separation," in *4th Joint Workshop on Hands-free Speech Communication Microphone Arrays*, Nancy, France, May 2014, (slides).

[28] Z. Rafii, B. Coover, and J. Han, "An Audio Fingerprinting System for Live Version Identification using Image Processing Techniques," in *39th IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, May 2014, (poster).

[29] Z. Rafii, F. G. Germain, D. L. Sun, and G. J. Mysore, "Combining Modeling of Singing Voice and Background Music for Automatic Separation of Musical Mixtures," in *14th International Society for Music Information Retrieval*, Curutiba, PR, Brazil, Nov. 2013, (poster).

[30] Z. Rafii and B. Pardo, "Online REPET-SIM for Real-time Speech Enhancement," in *38th IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada, May 2013, (poster).

[31] ——, "Music/Voice Separation using the Similarity Matrix," in *13th International Society for Music Information Retrieval*, Porto, Portugal, Oct. 2012, (slides).

[32] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive Filtering for Music/Voice Separation Exploiting the Repeating Musical Structure," in *37th IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012, (slides).

[33] M. Cartwright, Z. Rafii, J. Han, and B. Pardo, "Making Searchable Melodies: Human vs. Machine," in *3rd Human Computation Workshop*, San Francisco, CA, Aug. 2011, (poster).

[34] Z. Rafii and B. Pardo, "A Simple Music/Voice Separation Method based on the Extraction of the Repeating Musical Structure," in *36th IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011, (poster).

[35] ——, "Degenerate Unmixing Estimation Technique using the Constant Q Transform," in *36th IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, May 2011, (poster).

[36] ——, "Learning to control a Reverberator using Subjective Perceptual Descriptors," in *10th International Society for Music Information Retrieval*, Kobe, Japan, Oct. 2009, (poster).

5.4. Book Chapters

[37] B. Pardo, Z. Rafii, and Z. Duan, "Audio Source Separation in a Musical Context," in *Handbook of Systematic Musicology*, H. Springer Berlin, Ed. 2018.

[38] A. L. Zafar Rafii and B. Pardo, "REPET for Background/Foreground Separation in Audio," in *Blind Source Separation*, H. Springer Berlin, Ed. 2014.

5.5. Technical Reports

[39] Z. Rafii and B. Pardo, "A Digital Reverberator controlled through Measures of the Reverberation," Northwestern University, 2009.

5.6. Talks

[40] J. McDermott, B. Pardo, and Z. Rafii, *Leveraging Repetition to Parse the Auditory Scene*, 13th International Society for Music Information Retrieval, Porto, Portugal, 2012.

5.7. Data Sets

[41] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner. "MUSDB18-HQ – an uncompressed version of MUSDB18." (2019).

[42] ——, "The MUSDB18 corpus for music separation." (2019).