INSTRUMENTAL EXTRACTION IN HIP-HOP MUSIC EXPLOITING REPEATING STRUCTURE

Cole Peterson
University of Victoria
cpeterso@uvic.ca

George Tzanetakis
University of Victoria
gtzan@uvic.ca

ABSTRACT

Many blind source separation algorithms take advantage of repeating musical structures to better separate the voice from accompanying instrumental music. Hip-hop music is especially repetitive, often using identical loops under non-repeating rap lyrics. In this paper, we present a method for separating instrumental music and vocals in hip-hop using a two-step process. First similar sound segments are found using the track's magnitude spectrogram, and then an estimate of the repeated sounds within these segments is calculated. Our approach is based on the repeating pattern extraction technique (REPET) with modifications that improve its performance for hip-hop music. The most effective modification is finding the optimal alignment without constraining by tempo. However, this results in increased computation time compared to the original REPET technique. To evaluate performance, a new dataset of 50 popular hip-hop tracks and their instrumentals was assembled. Our proposed method was also tested on an existing dataset of 100 tracks containing a wider variety of music that has been used before for evaluation of existing methods. These experiments show that the proposed method can effectively separate full hip-hop tracks better than existing approaches, but this improvement is limited to hip-hop and other music with strong repeating structure.

1. INTRODUCTION

Hip-hop instrumentals have many uses, and this creates a demand for them. Instrumentals are reused by different rappers to release "mix-tapes", collections of songs less polished and less original than would be released on a proper album, they are used in battle rap competitions and by freestylers (rappers who improvise lyrics over a beat), they are used by remixers to blend two existing rap songs by putting the rap verse of one over the instrumental beat of another, and they are used by amateurs to perform karaoke. The demand for these instrumentals and the sampling culture of hip-hop means that many instrumentals are given an official release, thus being useful as a ground truth that is easier to come by for the evaluation of automated sound source separation. A new dataset of 50 songs, from varying artists and time periods, was created to evaluate the separation of hip-hop music, and guide the design and development of our algorithm.

The repetitive nature of hip-hop instrumentals (often composed of identical loops) also lends itself to sound source separation techniques that leverage repetition. The original REPET algorithm models music as a repeating background and non-repeating foreground, using repeating features of the sound to separate it. REPET segments the magnitude spectrogram into equally sized pieces, and then derives a "repeating mask" a soft-mask which estimates the repeating sounds in each segment and is used to filter every segment in the input sound. This method works well for short segments of music where the musical structure remains static, however it struggles to effectively separate full songs as it does not adapt over time to changes in the music. Several adaptations of the ideas in the original REPET algorithm have been produced to handle this common "verse-chorus" problem. Both repet_ada (Liutkus *et al.*) [10], and repet_seg (Rafii *et al.*) [13] look for repeating patterns locally, however, they do not take advantage of similar sounds which may appear far away from each other (e.g. instrumentals in the 1st verse and 3rd verse), nor does it exclude dissimilar sounds which appear close to each other (e.g. a verse right next to a chorus).

Adaptive REPET (repet_ada) [10] keeps track of a beat spectrogram, representing how the repeating period changes over the course of a song. It uses this beat spectrogram to find the repeating period p for each time slice, and then calculates the element-wise median of the two time slices $\pm p$ away from it for use in the repeating mask. Segmented REPET (repet_seg) [13] performs the original REPET algorithm on overlapping windows of a set size (with 10s shown to be optimal in [13]), overcoming original REPET's difficulty scaling to long songs by dividing it into smaller sounds. Both of these approaches are highly local, in repet_ada the repeating mask is determined by sounds p away, and in repet_seg the window size limits where repeating sounds can be found.

Like those adaptations, this paper presents a method that uses a *dynamic* repeating mask, which is able to change over the course of the song. The repeating mask is calculated by selecting a set number of segments of sound from anywhere in the track which are expected to contain the same or similar backgrounds. Additionally, while REPET and its variants use an element-wise median on the segments of the magnitude spectrogram, this paper suggests better results can be obtained by taking the minimum, or by using a frequency-specific neural-network function.

2. RELATED WORK

Audio source separation has been a topic of active research for several years. The Signal Separation Evaluation Campaign (SiSEC) has been instrumental in providing data sets and tasks for evaluating different algorithms [1, 20]. A large variety of algorithms have been proposed for this task [5, 6]. A somewhat recent overview can be found in Vincent [19]. In the most general setting no assumptions are made about the different sound sources in the mixture. A large number of work in audio source separation focuses on speech signal [8, 22] but more recently music signals have also been analyzed. Music provides a particularly interesting case as some sound sources can be considered more important than others (for example vocals) and assumptions about structure [2] and repetition can be used to improve the separation performance. The term informed sound source separation has been used to describe such algorithms [9]. Deep recurrent neural networks have been explored for signing voice separation [7]. Other approaches are based on non-negative matrix factorization [12] or utilize pitch and rhythmic information [14]. Our proposed approach belongs to a family of methods that leverage the strong repetition in musical signals that originate from the REPET (REpeating Pattern Extraction Technique) method [10, 13, 15, 16]. There are several approaches that have been proposed to evaluate the performance of blind and informed sound source separation algorithms [4, 18] originating from the audio source separation evaluation campaigns [20, 21].

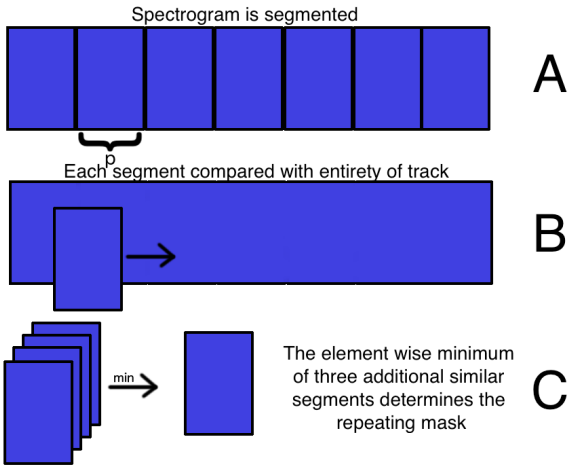
3. METHOD

Early stages of the algorithm match the REPET algorithm described in [15]. A spectrogram of the input signal, X, is calculated using a Short-Time Fourier Transform with a Hamming window and constant overlap. The magnitude spectrogram, V, is derived from X without the mirrored frequencies. The repeating period, p, is determined from autocorrelating each frequency bin of V^2, and taking the mean of the results.

3.1 Similar Segment Identification

A visual outline of the salient part of this algorithm can been seen in Figure 1. Like in the original REPET the magnitude spectrogram V is divided into segments of size p (step A in Figure 1), however we then compare each of these segments to every alignment possible over entirety of V(step B Figure 1). This allows us to find multiple similar sounding segments from anywhere in the track, while standard REPET uses its neighbouring segments (those which are offset by a multiple of p). Within local ranges the most similar segments of sound are often found offset by p, but can become misaligned over longer durations. We find better segments when not constraining the segments to a multiple of the repeating period: we search for the optimal alignment, using p only to determine the size of the segments. This provides for better results, but does come at a high computational cost, making this method slower than

Figure 1. Outline of the xcorr_min REPET algorithm



existing REPET methods. Given a p-length segment S, we identify 3 similar sounding segments by the process shown in Algorithm 1, which is essentially choosing the maximum 3 segments of a modified cross correlation between S and V.

Data: Magnitude Spectrogram V, Segment Magnitude Spectrogram S
Result: 3 similar sounding segments for each segment
$|s| = length(S)$;
$|v| = length(V)$;
for *i in* $(|v| - |s|)$ **do**
  similarity[i] = dot(S,min(S,V[i:i+p]));
  //dot(A,B) is the inner product of A and B
end
for *3 iterations* **do**
  j = argmax(similarity);
  add V[j:j+p] to similar segments of segment S;
  remove indexes in the range $j \pm /2$ from similarity;
end

Algorithm 1: Identification of 3 similar segments

3.2 Segment Operation

Once we have three similar sounding segments to our original segment, we compute a repeating mask, a soft mask used to filter the input signal. This repeating mask is entirely dynamic and will change for every segment, because each segment will have identified different segments which sound similar to it. The challenge then becomes determining the common sounds between the original segment and the three found elsewhere in V which sound similar. Many REPET-like algorithms take an element-wise median for every segment to calculate the repeating mask, although the mean has also been explored. We test two different operations on these segments, an element-wise minimum(shown in step C of Figure 1), and a learned neural network.

If you add a sound to x(t), you would expect the

magnitude spectrogram of the mixture to be greater than that of the original. This is because of the linear property of the Fourier transform, $\mathcal{F}\{x(t)\} = X(\omega)$ and $\mathcal{F}\{y_n(t)\} = Y_n(\omega)$: $\mathcal{F}\{x(t) + y_n(t)\} = X(\omega) + Y_n(\omega)$, so $|X(\omega) + Y(\omega)| > |X(\omega)|$

If you have multiple copies of x(t), with different sounds $y_n(t)$ attached, knowing that the original is less than each of the mixtures, you could take the minimum to recover the original sound.

$$\begin{aligned} x(t) \approx \mathcal{F}^{-1}\{min(&|X(\omega) + Y_1(\omega)|, \\ &|X(\omega) + Y_2(\omega)|, \\ &|X(\omega) + Y_3(\omega)|, \\ &|X(\omega) + Y_4(\omega)|)\} \end{aligned} \tag{1}$$

Although taking an element-wise minimum showed promising results, the frequency of the sounds we wish to remove (the rapped vocal sounds) lie within a predictable frequency range, and so using a operation accounts for this could be expected to bring better results, unlike the element-wise minimum which does not discriminate its output based on frequency bin. By training a neural network to operate on a given frequency bin, it might account for this and would also contain the potential to be more stable – if, for example, one of the selected segments contained a different underlying instrumental sound, a neural network might be able to account for it if the other two additional segments are acceptable, whereas an element-wise minimum will be highly sensitive to changes. While many papers have used deep neural networks (DNNs) on magnitude spectrograms to separate sound [7,8,11,17,22], many of these papers are trained to output hard masks, transforming the problem into a binary classification problem. Additionally, the data that is run through them is not processed to identify similar sounds before their use. In this paper, we explore training 1025 separate neural networks, one per frequency bin, to transform four sound segments to one soft mask. The networks were two layered feed-forward with 10 hidden neurons, the first layer using a sigmoid activation function and the second a linear function, and trained using Levenberg-Marquardt backpropagation.

Each neural network takes as input four values, one from each of the segments and outputs one value for use as a soft mask. The target values used in training are taken from the magnitude spectrogram of the actual instrumental, as this would be ideal output. The input values used in training come from similar segment processing described in Algorithm 1 done as pre-processing. Training was done on 10 randomly selected tracks from the dataset, and then 70% of the 4 value→1 value sets were used in training, 15% for test, and 15% for validation. 1025 different networks were trained, one for each frequency bin. For example, the neural network used for frequency bin 1 would be trained on data from the first frequency bin in the similar sounds, and the first frequency bin of the instrumental's magnitude spectrogram. When used to estimate the repeating mask, the four values from each segment identified by Algorithm 1 would be passed to the neural network for that frequency, and then the resulting value used as the repeating mask. The operation is element-wise and frequency dependent.

4. DATASETS

4.1 Hip-hop

A new dataset of 50 hip-hop tracks spread out over four decades of hip hop was compiled for evaluating and training. Full tracks were used wherever possible, occasionally clipping out "skit" intros. All tracks and their instrumentals had a sampling frequency of 44,100Hz, and were comprised of two stereo channels. As most released instrumentals were not properly aligned with their full tracks, this was done by hand in creating the dataset.

4.2 SiSEC MSD100 Dataset

The Signal Separation Evaluation Campaign has released the Mixing Secret Dataset, a collection of 100 songs split across genres for setting a benchmark for sound source separation. Following success on a strictly hip-hop dataset, the version of REPET was run on this general music dataset to gauge its performance on music in general. This dataset is also sampled at 44,100Hz, and has source files for drum, bass, vocals, and other. We considered the instrumental to be the combination of bass, drums, and other, and tested our output against that.

5. EVALUATION

Results were evaluated using Blind Source Separation Evaluation (BSS Eval) described in in [21]. This provides four objective measures of sound separation quality between a source and its estimate, a source to distortion ratio (SDR), source image to spatial distortion (ISR), source to interferences ratio (SIR), and source to artifact ratio (SAR). Definitions can be found in equations 2,3,4,and 5, further details can be found in [21]. A better separation will yield higher values. Figure 2 shows comparative results between REPET algorithms on the hip-hop dataset, Figure 3 shows the result on the MSD100 dataset. Table 1 shows how often each algorithm presents the best results for each of the BSS categories in the hip-hop dataset.

Table 1. Number of songs each algorithm gives the best result for each BSS measure on the hip-hop dataset.

	xcorr_min	repet_ada	repet_seg
SDR	**29**	1	20
ISR	**32**	6	12
SIR	**19**	14	17
SAR	**33**	1	16

$$SDR := 10log_{10}\frac{|s_{target}|^2}{|e_{interf} + e_{noise} + e_{artif}|^2} \tag{2}$$

$$SIR := 10log_{10}\frac{|s_{target}|^2}{|e_{interf}|^2} \quad (3)$$

$$ISR := 10log_{10}\frac{|s_{target}|^2}{|e_{noise}|^2} \quad (4)$$

$$SAR := 10log_{10}\frac{|s_{target} + e_{interf} + e_{noise}|^2}{|e_{artif}|^2} \quad (5)$$

Figure 2. SDR performance of different methods on hip-hop dataset.

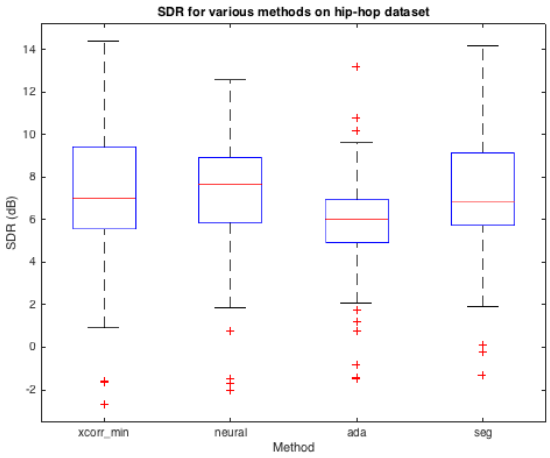


Figure 3. SDR performance of different methods on MSD100 dataset.

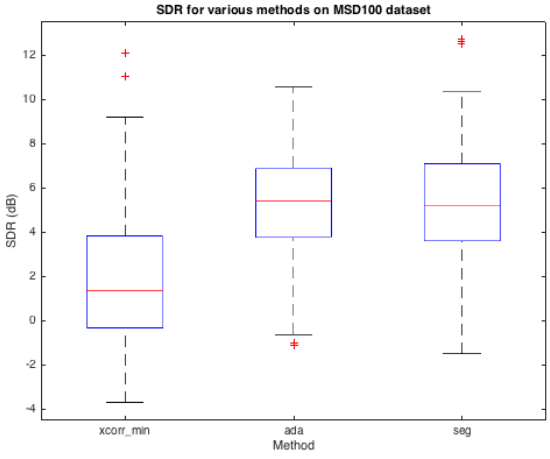


Table 2. Mean and standard deviation performance for each algorithm on the hip-hop dataset

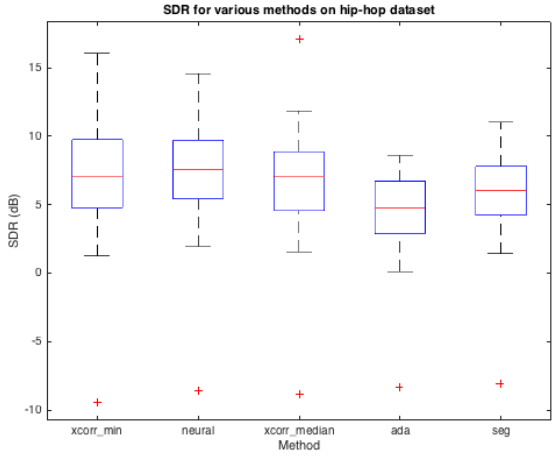
	SDR	ISR	SIR	SAR
xcorr_min	7.1 ± 3.8	**13.8 ± 5.8**	23.5 ± 5.0	8.7 ± 4.3
xcorr_neu	**7.1 ± 3.3**	12.9 ± 5.0	23.8 ± 4.3	**8.8 ± 4.0**
repet_ada	5.6 ± 2.9	13.1 ± 5.2	26.2 ± 5.1	6.2 ± 3.5
repet_seg	6.8 ± 3.0	13.3 ± 5.3	**26.5 ± 5.1**	7.7 ± 3.6

Table 3. Mean and standard deviation performance for each algorithm on the MSD100 dataset

	SDR	ISR	SIR	SAR
xcorr_min	1.9 ± 3.1	9.5 ± 3.8	30.2 ± 35.2	5.2 ± 2.5
repet_ada	5.2 ± 2.6	**14.6 ± 3.4**	**36.5 ± 32.7**	7.0 ± 2.2
repet_seg	**5.3 ± 2.9**	14.1 ± 4.0	36.4 ± 32.7	**7.3 ± 2.4**

Results within the hip-hop dataset show that both the neural and xcorr_min versions of repet presented in this paper outperform repet_ada and repet_seg at the mean and different percentiles. However, we find that the subjective quality of the instrumentals extracted by xcorr_min greatly exceed the results in any of the other methods, which is not adequately captured in the bss_eval numbers. Output from the neural method is often accompanied with noisy artifacts.

Figure 4. SDR performance of different methods hip-hop dataset



We also tested the xcorr front-end with an element-wise median to isolate the effect that segment selection would have. This produced better results than either repet_ada or repet_sim, but did not perform as well as xcorr_min. Those results can be see in Figure 4.

6. DISCUSSION AND FUTURE WORK

In the process of experimenting with hip-hop tracks we discovered two limitations of the proposed method: 1) muted end loops, 2) stereo signals. A common technique in hip-hop production is to mute the instrumental at the end of a line to place emphasis on the lyrics. In the data set, this can be seen in "All Caps" (allcaps.wav), and "Ain't No Half Steppin'" (half.wav), among others. However, if one of these segments is chosen for use with a segment where the end of the loop is not muted, this will over restrict the repeating mask, and only allow the frequencies that are present in the vocal. For example, consider a loop that repeats four times, and on the fourth, the end is muted. The algorithm put forward in this paper will use a repeating mask that eliminates instrumental frequencies at the end of

the loop, when the right choice would be to keep those frequencies in the mask, as when considering the output for the fourth loop only the frequencies present in the segment will be output and not those in the mask (Stage 3 of REPET).

REPET treats the two channels independent of one another, but this neglects additional and potentially valuable information present in the other channel. Any sounds that change channels will be thrown out by the repeating mask. It should also be noted that stereo features could be used to improve the separation, as vocals are usually panned in the center, whereas instrumental sounds are often panned more to the right or left channels. In some cases, subtracting the one channel from another in the time-domain can result in surprisingly good separation.

In the future we plan to add automatic detection of muted-end loops and incorporate stereo panning information in the creation of the mask to improve vocal separation without making the strong assumption that vocals are always centered. When listening to the separation results it is easy to identify which algorithm is used. Also frequently the perceptual quality of the separated vocals and instrumentals not perfectly correlated with the SDR (or other objective metric) value. This is something that is known in the audio source separation community and has led to the development of the Perceptual Audio Source Separation toolkit [3] which we plan to use.

From our experiments it is clear that the proposed method performs better in Hip Hop music but not as well in the more general MSD100 data set. We plan to investigate the automatic detection of which variant of the algorithm is more appropriate as a way to improve the results. To support reproducibility the code and associated data set described in this paper are available by request from the authors.

7. REFERENCES

[1] Shoko Araki, Francesco Nesta, Emmanuel Vincent, Zbynek Koldovsky, Guido Nolte, Andreas Ziehe, and Alexis Benichoux. The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Audio source separation -. In *10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pages 414–422, Tel Aviv, Israel, March 2012.

[2] Roger B. Dannenberg and Masataka Goto. Music structure analysis from acoustic signals. *Handbook of Signal Processing in Acoustics*, pages 305–331, 2008.

[3] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. The PEASS Toolkit - Perceptual Evaluation methods for Audio Source Separation. 9th Int. Conf. on Latent Variable Analysis and Signal Separation, September 2010.

[4] Valentin Emiya, Emmanuel Vincent, Niklas Harlander, and Volker Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 19(7):2046–2057, September 2011.

[5] Cédric Févotte. *Bayesian Audio Source Separation*, pages 305–335. Springer Netherlands, Dordrecht, 2007.

[6] D. B. Haddad, Diego Barreto Haddad, M. R. Petraglia, Mariane Rembold Petraglia, P. B. Batalheiro, and Paulo Bulkool Batalheiro. Performance evaluation of two semi-blind source separation methods. In *2008 IEEE 9th Workshop on Signal Processing Advances in Wireless Communications*, pages 231–235. IEEE, 2008.

[7] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Singing-voice separation using deep recurrent neural networks. In *Music Information Retrieval Exchange (MIREX), 2014*.

[8] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Deep learning for monaural speech separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1562–1566. IEEE, 2014.

[9] Antoine Liutkus, Jean-Louis Durrieu, Laurent Daudet, and Gaël Richard. An overview of informed audio source separation. In *WIAMIS*, pages 1–4, Paris, France, 2013.

[10] Antoine Liutkus, Zafar Rafii, Roland Badeau, Bryan Pardo, and Gaël Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 53–56. IEEE, 2012.

[11] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. *Multichannel audio source separation with deep neural networks*. PhD thesis, INRIA, 2015.

[12] Alexey Ozerov, Emmanuel Vincent, and Frédéric Bimbot. A General Flexible Framework for the Handling of Prior Information in Audio Source Separation. *IEEE Transactions on Audio, Speech and Language Processing*, 20(4):1118 – 1133, May 2012. 16.

[13] 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*, 21(1):73–84, 2013.

[14] Zafar Rafii, Zhiyao Duan, and Bryan Pardo. Combining rhythm-based and pitch-based methods for background and melody separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1884–1893, 2014.

[15] Zafar Rafii and Bryan Pardo. A simple music/voice separation method based on the extraction of the repeating musical structure. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 221–224. IEEE, 2011.

[16] Zafar Rafii and Bryan Pardo. Online repet-sim for real-time speech enhancement. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 848–852. IEEE, 2013.

[17] Stefan Uhlich, Franck Giron, and Yuki Mitsufuji. Deep neural network based instrument extraction from music. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2135–2139. IEEE, 2015.

[18] E. Vincent, R. Gribonval, and C. Fevotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4):1462–1469, 2006.

[19] Emmanuel Vincent. Advances in audio source seperation and multisource audio content retrieval. In *SPIE Defense, Security, and Sensing*, pages 840109–840109. International Society for Optics and Photonics, 2012.

[20] Emmanuel Vincent, Shoko Araki, Fabian Theis, Guido Nolte, Pau Bofill, Hiroshi Sawada, Alexey Ozerov, Vikrham Gowreesunker, Dominik Lutter, and Ngoc Q. K. Duong. The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936, 2012.

[21] Emmanuel Vincent, Hiroshi Sawada, Pau Bofill, Shoji Makino, and Justinian P. Rosca. *First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results*, volume 4666, pages 552–559. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[22] Felix Weninger, John R Hershey, Jonathan Le Roux, and Bjorn Schuller. Discriminatively trained recurrent neural networks for single-channel speech separation. In *Signal and Information Processing (GlobalSIP), 2014 IEEE Global Conference on*, pages 577–581. IEEE, 2014.