Voice & Music Pattern Extraction: A Review

Pooja Gautam[1] and B S Kaushik[2]

[1]Electronics & Telecommunication Department
RCET, Bhilai,
Bhilai (C.G.) India
pooja0309pari@gmail.com

[2]Electrical & Instrumentation Department
Bhilai (C.G.) India
Bhuneshwersingh.kaushik@gmail.com

Abstract: ***This paper presents a review of three most popular methods of separation are based on "repeating pattern extraction technique (REPET)", "Pitch based method"& "Hybrid based method" techniques involved in the separation of voice & music from a mixture (Song). The comparison has been made on the basis of SIR, SAR & SDR. On comparing this method, it was found that Hybrid method of series combination of Pitch & REPET gives better performance then rest of the methods.***

Keywords*:* REPET, Pitch, Signal to interference ratio(SIR), Signal to artifacts ratio(SAR), Signal to Distortion ratio (SDR).

1. INTRODUCTION

Musical works are often composed of two components: the background (typically the musical accompaniment), which generally exhibits a strong repeating structure with distinctive repeating time, and the melody (typically the singing voice), which generally exhibit the strong harmonic structure with a distinctive pitch contour. Drawing from the findings in cognitive psychology, we propose to investigate the combination of simple of two approaches for separating those two components: a REPET method that focuses on background extraction via a rhythmic mask derived from identifying the repeating time elements in mixture and a pitch-based method that focuses on extracting the melody via a harmonic mask that is derived from identifying the predominant contours of pitch in the mixture. Evaluation on a data set of song clips showed that combining of such of the two contrasting yet complementary methods can help to improve the separation performance from the point of view of both of the components compared with using only one of those methods, and also compared with the two other state-of the- art approaches**.** An instrumental track containing only the instruments for researcher's application that includes:
Studying MIR (Music Information Retrieval), It could be used in Active Noise Control for removing periodic interferences, Applications includes: Cancelling periodic interferences in electrocardiography (e.g., the power-line interference) & In speech signal (e.g., pilot communicating by radio from an aircraft). Also can be applied for periodic interferences removal, This is a problem of great interest for both entertainment industry & researchers. For this project, I compared the performance / Merits & Demerits of different algorithms which can be used for music/voice separation.
The organisation of the paper is as follows literature survey is given in section 2, section 3 gives conclusion of literature review. Section 4 gives an idea about problem related to voice & music separation. Section 5 Three different methods related to separations are being discussed in this section. Section 6 gives the result of various methodologies which are reviewed and section 7 which concludes the paper.

2. LITERATURE REVIEW

Hsu et al. (2012) proposed a pitch based separation system. A trend estimation algorithm first estimates the pitch ranges of singing voice. The estimated trend is then incorporated in tandem algorithm to acquire the initial estimate of the singing pitch. Singing voice is separated according to the initially estimated pitch. The above two stages, i.e., pitch determination and voice separation iterate until its convergence. A post processing stage is introduced to deal with sequential grouping problem, i.e., deciding which of those pitch contours belong to the target, an issue unaddressed in the original tandem algorithm. Finally, singing voice detection is performed to discard the non vocal parts of the separated singing voice. Furthermore, the boundary for upper pitch of singing can be as high as 1400 Hz for soprano singers while pitch range of normal speech is between 80 and 500 Hz. The differences make the separation of singing voices and the music accompaniment potentially more challenging.

RAFI and Pardo (2012) proposed a method on the assumption that Repetition is a fundamental element in the generation and perceiving structure in the music. This method separates the musical background from the vocal foreground in the mixture, Instead of looking for periodicities; this method uses a similarity matrix to identify the repeating. It then calculates the repeating spectrogram model using the median and extracts the repeating patterns using a time-frequency masking.
Proposed system doesn't supports for the small rhythmic patterns, but the rhythmic patterns are essential for the balance of music, and can be a way to identify a song.

RAFI and Pardo (2011) proposed new method which also uses the repetition property of music in song, and separates the voice & music. In this method first, the period of the repeating structure is found. Then spectrogram is segmented at the period boundaries and thus segments are averaged to create a repeating segment model. Finally, each of the time-frequency bins in the segment is compared to the model, and mixture is partitioned using binary time-frequency masking

Cases where repetitions also happen intermittently or without a fixed period

RAFI and Pardo (2013) proposed new method unlike above previous approaches; this method does not depend on particular features, does not only rely on complex frameworks but also does not requires prior training. Because it is only based on the self-similarity, the existing method could potentially work on any audio, as long as there are repeating structures. It has therefore the advantage that the method is simple, fast, blind, and also completely automatable.

The basic idea is to: A. identifies the periodically repeating segments, B. Compare them to a repeating segment model, and C. Extract the repeating patterns via time-frequency masking.

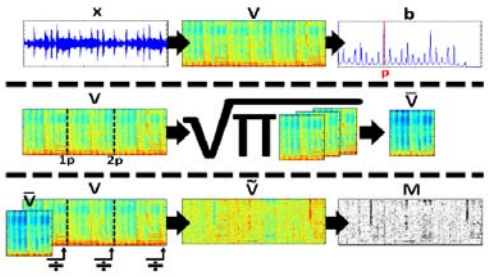


Figure.1: REPET Clearly states the procedure

RAFI & Duan (2014) proposed hybrid methods with two different combination of REPET & Pitch based method:

In Parallel combination, from given a mixture spectrogram, REPET derives a back ground mask and the complementary melody mask and Pitch derives a melody mask and the complementary background mask. The final background mask and the final melody mask are then derived by weighting and Wiener filtering.

In series combination from given a mixture spectrogram, REPET first derives a background mask and the complementary melody mask. Given the melody mask, Pitch then derives a refined melody mask and a complementary "leftover"mask . The final background mask and the final melody mask are then derived by weighting and Wiener filtering (WF) the masks.

3. CONCLUSION OF LITERATURE REVIEW

Number of methods applied for separating the repeating "background" from that of the non-repeating "foreground" in a mixture for a monaural singing voice separation, and the existing methods can be generally classified as these three categories below depending on the underlying methodologies: spectrogram factorization methods, model-based methods, and pitch-based methods.

3.1 Spectrogram factorization/ REPET: In this existing method of Music & voice separation the Music accompaniment can be assumed to be in a low-rank subspace, on the other hand, singing voice can be regarded as relatively sparse within songs, also the repetition property of music is utilised to separate the music and voice based on this assumption different methods like RPCA /REPET is used for solving the underlying low-rank and sparse matrices.

3.2 Pitch-based methods: In this method the property of voice & music that it is having different pitch ranges, range of normal speech is between 80 and 500 Hz and pitch range of music is higher than 500Hz. Initially it estimates the pitch range of singing voice and then separated according to the estimated pitch. The above two stages pitch determination and voice separation then iterate until convergence.

3.3 Hybrid model: Hybrid methods model by combining different methods. Cobos et al. used a panning-based method and a pitch-based method. Virtanen et al. used a pitch-based method to first identify the vocal segments of the melody and an adaptation-based method with NMF to then learn a model from the non-vocal segments for the background. Wang et al. used a pitch-based method and an NMF-based method with a source-filter model. FitzGerald used a repetition-based method to first estimate the background and a panning-based method to then refine background and melody. Rafii et al. used an NMF-based method to first learn a model for the melody and a repetition-based method to then refine the background.

4. PROBLEM IDENTIFICATION

Pitch based Method best suited for non repeating pattern extraction but it having limitations that we have to find out the exact pitch value of singing voice and it's a difficult task to clearly differentiate the singing voice & instrumental pitch ranges. But having efficient property of removal of odd pitch spectrum.

REPET method also separates the repeating pattern and gives the higher values of SDR & GNSDR compared to all other known methods as shown in Fig.3.1, SDR values for obtained by REPET shown in fig.3.2. The REpeating Pattern Extraction Technique (REPET) separates the repeating audio signal from the non-repeating audio signal in a mixture. The basic idea is to identify the periodically repeating segments in the audio, compare them to the repeating segment model derived from them, and extract the repeating patterns via time-frequency masking.

Method gives best result for separation of repeating beat structure, but fails to separate the non repeating beats and the non repeating beats of musical instruments as it is lying in voice signal.

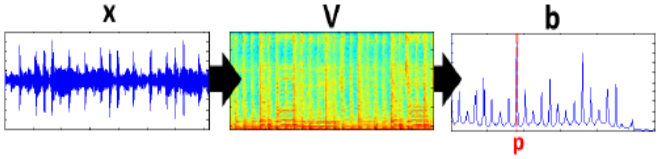
5. METHODOLOGIES

A hybrid method for Voice & Music separation based on REPET and Pitch based will be used. In first part of project we will apply the REPET method on given input mixture which will separate out the repeating & non repeating part. In Second part we will apply pitch based method on the output to separate out the higher pitch value signals which are non repeating beats.

5.1 REpeating Pattern Extraction Technique (REPET): Repetition is a core principle in music. Many musical pieces are thus characterized by an underlying

repeating 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.

5.1.1 Identify the periodically repeating segments,

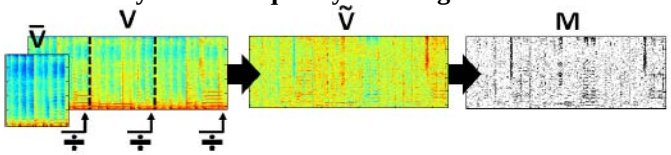


Periodicities in a signal can be found by using autocorrelation, which measures the similarity between a segment and a lagged version of itself over a successive time intervals.
Given a mixture signal x, Method first calculate its Short-Time Fourier Transform X, using half-overlapping Hamming windows of N samples. Then derive the magnitude spectrogram V by taking the absolute value of elements of X, after discarding the symmetric part, while keeping the DC component. Then compute the autocorrelation of each row of power spectrogram V2 (element-wise square of V) and obtain the matrix B. Method use V2 to emphasize the appearance of the peaks of periodicity in B. If the mixture signal x is stereo, V2 is averaged over the channels. And 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 (lag 0).

5.1.2 Repeating Segment Model:

After estimating the period p of the repeating musical structure, method uses it to evenly segment the spectrogram V into segments of length p. Then compute a mean repeating segment V over r segments of V, which can be thought of as the repeating segment model. The idea is that time-frequency bins comprising of the repeating patterns would have similar values at each period, and would also be similar to the repeating segment model. Experiments showed that the geometric mean leads to a better extraction of repeating musical structures than the arithmetic mean.

5.1.3. Binary Time-Frequency Masking:



After computing the mean repeating segment V-,method divide each time-frequency bin in each segment of the spectrogram V by the corresponding bin in V- . Then take the absolute value of logarithm of each bin to get a modified spectrogram ~V and furthermore the repeating musical structure generally involves some variations. Therefore, method introduce a tolerance t when creating the binary time frequency mask M. experiments show that a tolerance of t = 1 gives good separation results, both for music and voice.

Once the binary time-frequency mask M is computed, it is symmetrized and applied to the STFT X of the mixture x to get the STFT of the music and the STFT of the voice The estimated music signal and voice are finally obtained by inverting their corresponding STFTs into the time domain.

5.2 A Tandem Algorithm for Singing Pitch Extraction: The pitch based system is illustrated in Fig. 4.1. A trend estimation algorithm first estimates the pitch range of the singing voice. The estimated trend is then incorporated in the tandem algorithm to acquire the initial estimates of the singing pitch. Singing voice is then separated according to the initially estimated pitch. The above two stages, i.e., pitch determination and voice separation then iterate until convergence. A post processing stage is introduced to deal with those "sequential grouping" problems, i.e., deciding which pitch contours belong to the target, an issue unaddressed in the original tandem algorithm. Finally, singing voice detection is performed to discard the non vocal part

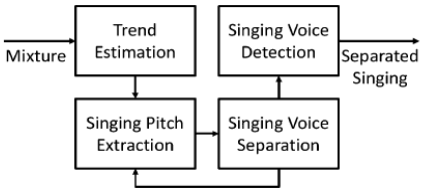


Figure.2 A tandem algorithm for singing pitch extraction

5.2.1 Trend Estimation: First, the singing voice is enhanced by considering temporal and spectral smoothness. As the fundamental frequency of the singing voice tends to be smooth across time, we bound the vocal in a series of time–frequency blocks. The T-F blocks give rough pitch ranges along time which are much narrower than the possible pitch range.

5.2.2 Pitch Range Estimation: The main objective of this stage is to find a sequence of relatively tight pitch ranges where the singing voices are present. The main idea to achieve this goal is to remove unreliable peaks not originating from periodic sounds and then higher harmonics of the singing voice. The remaining peaks approximate fundamentals and we estimate the range by bounding the peaks in a sequence of T-F blocks.

5.3 Hybrid method: A hybrid method for Voice & Music separation based on REPET and Pitch based will be used, RAFI & Duan (2014) proposed hybrid methods with two different combinations of REPET based & Pitch based methods:

In Parallel combination, from given a mixture spectrogram, REPET derives a back ground mask and the complementary melody mask and Pitch derives a melody mask and the complementary background mask. The final background mask and the final melody mask are then derived by weighting and Wiener filtering.

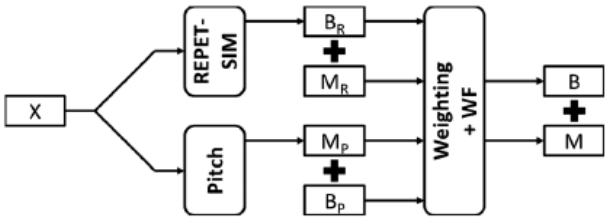


Fig. 5.3.1 Parallel Hybrid model

In series combination from given a mixture spectrogram, REPET first derives a background mask and the complementary melody mask. Given the melody mask , Pitch then derives a refined melody mask and a complementary "leftover" mask. The final background mask & the final melody mask are then derived by weighting and Wiener filtering (WF) the masks.

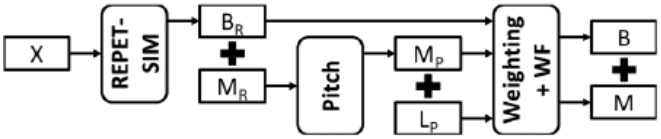


Figure.3: Series Hybrid model

Now as we earlier experience that the voice part contains some high pitch value Beats, to remove that beats pitch is estimated and to reach to the exact values of Beats process it repeated till the target pitch will be removed from the voice.

6. RESULT

The separation performance evaluated by employing the BSS EVAL toolbox. The toolbox proposes a set of now widely adopted measures that intend to quantify the quality of the separation between the source and its corresponding estimate: Source-to-Distortion Ratio, Sources-to-Interferences Ratio, & Sources-to-Artifacts Ratio.

$$SDR = 10\log_{10}\left(\frac{\|s_{target}\|^2}{\|e_{interf} + e_{artif}\|^2}\right)$$

$$SIR = 10\log_{10}\left(\frac{\|s_{target}\|^2}{\|e_{interf}\|^2}\right)$$

$$SAR = 10\log_{10}\left(\frac{\|s_{target} + e_{interf}\|^2}{\|e_{artif}\|^2}\right).$$

$$\hat{s}(t) = s_{target}(t) + e_{interf}(t) + e_{noise}(t) + e_{artif}(t)$$

Where **Starget** is an allowed distortion of source S and **eintrf**, **enoise** and **eartif** represents respectively the interferences of the unwanted sources, the perturbation noise and artifacts introduced by separation performance.

Higher values of SDR, SIR, and SAR suggest better separation performance.

Table 1. Comparison of performance of various methods

	For Background/Music		
Methods	SIR	SAR	SDR
	DB		
REPET	-8.5	-6.5	-7.8
PITCH	-10	-9.7	-10
Parallel Hybrid	-8.9	-5.4	-7.4
Series Hybrid	-10	-4.1	-7.9

Table 2. Comparison of performance of various methods

Methods	**For foreground/Voice**		
	SIR	SAR	SDR
	DB		
REPET	-15	-3.1	-9.9
PITCH	-11	-9.7	-11
Parallel Hybrid	-13	-3.2	-9
Series Hybrid	-13	-3.2	-8.9

7. CONCLUSION

The SIR, SAR & SDR obtained for background for various methods is given in table 1, the SDR is an overall performance measure that combines degree of source separation (SIR) with quality of the resulting signals (SAR), SDR by using parallel hybrid method it is found -7.4 DB which is highest.

The SIR, SAR & SDR obtained for foreground for various methods is given in table 2. the SDR for foreground is found -9 for parallel hybrid method & -8.9 for series hybrid method which is highest from the two other methods.

Hybrid compression gives better performance than rest of the methods. The hybrid method is combination of REPET & Pitch based method.

References

[1] Zafar Rafii , Zhiyao Duan, and Bryan Pardo (2014), 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, december 2014, pp. 1884-1893*

[2] Z. Duan, J. Han And B. Pardo (2013),Multi-pitch Streaming of Harmonic Sound Mixtures. *manuscript for ieee trans. audio, speech and language processing. pp. 1-13*

[3] Po-Sen Huang, Scott Deeann Chen (2012). Singing-voice separation from monaural recordings using robust

principal component analysis. *Paris Smaragdis, Mark Hasegawa-Johnson IEEE. ICASSP,* pp. 57-60.

[4] Hsu Chao-Ling, Wang D., Roger J. Jyh-Shing, and Hu K. (2012). A Tandem Algorithm for Singing Pitch Extraction and Voice Separation from Music. *IEEE transactions on audio, speech, and language processing, vol. 20, no.5,* pp. 1482-1491.

[5] Zafar RAFII, Bryan Pardo (2011). A simple music/voice separation method based on the extraction of the repeating musical structure. *36th International Conference on Acoustics, Speech and Signal Processing, ICASSP* , pp.1-4

[6] Zafar Rafii, Student Member, IEEE, and Bryan Pardo, Member, IEEE (2013). 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.

[7] Paris Smaragdis and Judith C. Brown (2003). Non-Negative Matrix Factorization for Polyphonic Music Transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustic,* pp.177-180.