REpeating Pattern Extraction Technique (REPET), a Novel and Simple Approach for Separating the Repeating "Background" from the Non-repeating "Foreground" in a Mixture

Priyanka B. Sabale
M. E. Scholar,
Department of Electronics and Telecommunication,
N. K. Orchid College of Engineering and Technology, Solapur, Maharashtra

Prabhavati D. Bahirgonde
Associate Professor
Department of Electronics and Telecommunication
N. K. Orchid College of Engineering and Technology, Solapur, Maharashtra

ABSTRACT
As far as music is concerned the repetition forms the core. The repeating structure forms the characteristic of the musical piece and superimposed by various elements especially in the case of pop music. Proposed work concentrates on the Indian folk music and music voice separation from audio. The current available separation techniques are well suited for some typical music like REPET goes well with pop music as it has a constant repeating. Thus the study tries to present REPET for separating the repeating background from non repeating forebackground. The process starts with identifying the segments that repeats periodically, further comparison of this with the repeating segment derived model and then extraction of the repeating patterns using the time frequency masking. The Linear Predictive Coding **(**LPC) technique is used by most of the voice recognition devices in order to get input speech signal spectral information. LPC is used to generate observation vectors that are used by voice recognizers. On the basis of the parameters like Signal to Noise Ratio (SNR) the two techniques LPC and REPET are compared in the present work.

Keywords
REPET, LPC, MIR

1. INTRODUCTION
As per the analysis of audio and speech recognition is concerned, voice of singer and music separation forms the recent topics of research. In a field of music the above topics have good number of applications like music structure determination, recognition of lyrics and recognition of singer. A considerable number of studies dealing with the voice and background separation have carried out but very less is concerned with the signing voice in particular. In this work REPET is used for developing a novel methodology that can be used for the betterment of the vocal and non vocal separation process [1]. The rectification of the problems identified in the REPET is carried during the improvement of repeating mask that can be used for extracting the non vocal part in the audio. REPET being an independent in nature so it is preferred in the current work over other techniques.

The methods that exist for the separation of vocal and non vocal part are based on assumptions.

The art of music is the effect of the instruments being played and the voice of the human having repeating/non repeating pattern. Music forms an inseparable part of life in the various cultures. As far as security and authentication n of devices [2], tracking of emotions based on voice [3] is concerned analysis of speech plays a vital role.

The human voice has various components of frequency with varying amplitudes and depends on the person who is generating it. The audible range for human being is around 100-3200 Hz . As far as male is concerned it is 70-200Hz and for females it is 140-400 Hz. But voice found in signing has a wide range of frequency that can go utpo kHz. The pronunciation of words and sounds has an effect on the frequency of the person during speech or singing. This plays an important role in the differentiating the voice of different speakers [4]. Depending on the frequency content the research on analysis of voice is carried out by exploring various features of sound. Songs have a mixture of voice as well as music created by instruments. As compared to humans the musical instruments also develops sound at different amplitudes and frequencies.

Features like Mel-frequency cepstral coefficients (MFCCs) are used in detecting the vocal segments and then separation methods like Matrix Factorization [5], pitch-based inference [6],[7], or adaptive Bayesian modeling [8] are used in Music/voice separation systems. As mentioned earlier REPET does not rely on complex frameworks and does not depend on particular features, and does not require prior training. It is a method based on self similarity and can be used for any audio having repeating structure. Thus it can be noted that it is fast, simple, blind and automatable.

Need of Study
Identification of singer or instrument, extraction of melody and analysis of audio content are the matter of interest in separating voice and music and as gained importance in currently. This also helps people to record their voice on an music piece or wish to sing using the music without the original vocal. Separation of the vocal and the music woks upon the original mixture and provides the music alone.

2. GAPS IDENTIFIED
For the separation of music and the singing voice researchers have proposed different methods and algorithms. The

differentiation of the singing voice pitch and the instrumental pitch range is difficult; this forms the limitation of the Pitch based method which is best for extraction of non-repeating pattern. Model based method works well for extraction of repeating patterns as they have efficiency in removing odd pitch. But this method needs proper training. On the other hand REPET method discriminates the repeating pattern and separates the repeating signal of voice from non repeating one in a mixture. The basic idea is to identify the segments that repeat periodically in an audio, comparison of the same with the repeating segment derived model, further extraction of the repeating patterns using time frequency masking. This method is best suited for separating repeating structure; on the other hand it does not work for non repeating beats and the these non-repeating beats of musical instruments as it is lying in voice signal.

3. REPET METHODOLOGY

The overall REPET method can be summarized in three stages, namely (I) identification of the repeating period, (II) repeating segment modeling, and (III) repeating patterns extraction. In this work, this method was chosen over other methods presented in the literature because many musical works indeed include a repeating background (background music) overlaid on the non repeating foreground (singing voice). Moreover, repetition was recently used for source separation in studies of psychoacoustics. Repetition forms the basis of research work in different fields involving speech recognition and language detection, and also in MIR.

The idea of the REPET method is to identify the repeating structure in the audio and use it to model a repeating mask. The mask can then be compared to the mixture signal to extract the repeating background. The REPET method explicitly assumes that the music work is composed of repeating patterns.

A. Repeating Period Identification

With time interval of 0.04 seconds, 2048 samples and frequency of 44100 HZ, the Short-Time Fourier transform of mixture signal in MATLAB is calculated; the mixture spectrogram for the whole song can be obtained as shown in figure 1. Using the autocorrelation on mixture spectrogram, that is, comparing the segment and its lagged version over successive time interval to measure the similarity in the segment. The rows of mixture spectrogram are slide and calculate the autocorrelation of each row to get a matrix B. After that, the mean value for each row of the matrix B to get the beat spectrum is computed. After normalization using the first term of beat spectrum b, the final beat spectrum is achieved. If a mixture contains the repeating structure, there will be several peaks occur periodically in the beat spectrum. The basic idea is that the time between two peaks that occur periodically in the beat spectrum is the repeating period needed. Figure 2 shows a beat spectrum of one of the experiment of a song. The repeating period clearly is seen from the beat spectrum.

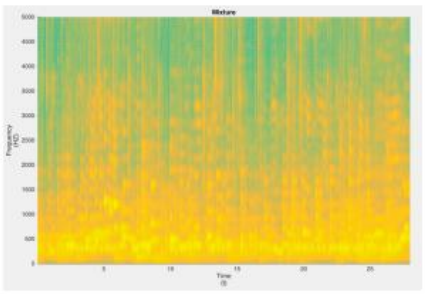

Figure 1 Mixture Spectrogram

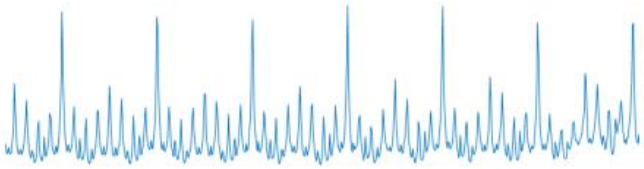

Figure 2 Beat Spectrum

B. Repeating Segment Modeling

After obtaining the repeating period, the repeating period to evenly time segment the mixture spectrogram into several segments of a length of the repeating period can be used which is shown in Figure 3. Then, in order to get the repeating segment, the element-wise median of time-frequency bin of each segment of the mixture spectrogram is calculated and takes this median as the repeating segment model. Since, the mixture spectrogram is segmented according to the repeating period, the median of each segments of the mixture spectrogram should be able to capture the repeating pattern of music background and remove the non-repeating singing voice foreground without the impact of outliers.

Dividing the spectrogram (V), into segments (r) of length (p) is the first step for calculating the repeating segment model.

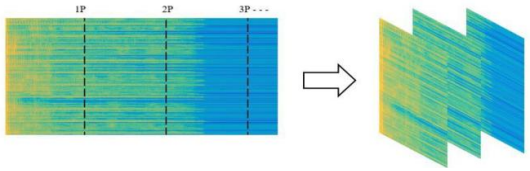

Figure 3: Segmentation of magnitude spectrogram V into r segments

The repeating segment is calculated as the element-wise median of the segments r of V. With the help of equation (1) the calculations for the model of segment being repeated can be stated mathematically.

$$S(i, j) = \text{median}\{V(i, l+(k-1)p)\} \quad (1)$$

Where, frequency index (i) = 1 . . . n and time index l = 1 . . . p here p stands for repeating period length.

The time frequency representation of non repeating voice as compared to music (repeating) is scattered and varied, hence the median of segments r is taken for repeating segment modeling. Every segment of the spectrogram V presents repeating structure for audio and few components that are non repeating, that might be the singers voice.

The non repeating part of audio is eliminated while the repeating structure is retained by considering the median of all the segments. When mean is considered, there are some shadows of non repeating elements found which are absent if median is considered.

4. REPEATING PATTERNS EXTRACTION

Once the model S of repeating segment is obtained it is repeated in order to match spectrogram length. Then the repeating spectrogram is obtained that is minimum in context of element wise among the latest model S of repeating segment and the respective segment of magnitude spectrogram

The repeating spectrogram model is calculated using the equation 2

$$W(I,l+(k-1)p) = \min \{S(i,l), V(i,l+(k-1)p)\} \quad (2)$$

Using the spectrogram V to normalize the repeating spectrogram model soft mask M is calculated. At a period of p in the spectrogram V the time frequency bins repeat and have values near to 1 but are supposed to have values close to 0. Thus normalization of W with respect to V gives values that are likely to repeat for each samples of p. Finally the soft mask M is applied to signal Short-time Fourier transform (STFT) and inverse STFT to the result to unpack the frequency bins to audio samples. The last step of the REPET is shown in figure 4

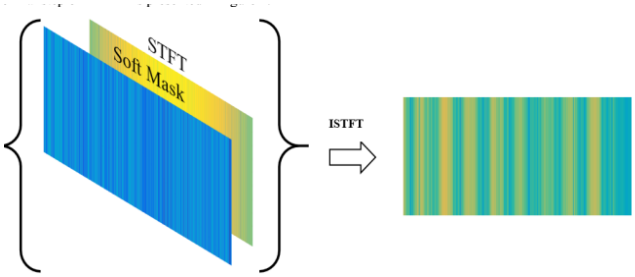

Figure 4: Estimation of background and unwrapping of signal using ISTFT

A. Sampling

It is important ot note that as far as principle is concerned MATLAB programming using the vectors of real numbers can be used to represent the digital audio signals as is the case with discrete-time signal. As far as the discrete time term refers that in nature time runs n continuum, whereas in the digital world real world signal samples drawn on discrete-time instances can be manipulated. The said process is called as sampling, which is the first step in creating the digital signals from its real world counterpart. The second step of sample quantization is ignored for the moment. For example, say at time 0 the first sample is recorded (commencing of measurement), the second at 0.001s, and the third at 0.002s and further. Here the instances of time are equidistant, computing the difference between any two consecutive instances of time its results into Sampling Period Ts= 0.001s.

It can be stated that for every Ts second one sample is drawn. The celebrated sampling frequency is the inverse of TS. Here the unit for sampling frequency is Hz and is equal to Fs = 1000 Hz i.e. 1000 samples of the real-world signal are taken every second.

A matter of concern is how high the sampling frequency should be or how short the sampling period should be. To sample a continuous time signal successfully the sampling frequency needs to be set equal to at least twice the maximum signal of frequency. This is known as the Nyquist rate and it ensures that aliasing is avoided which is an undesirable phenomenon as it introduces distortion i.e. quality of resulting audio is reduced.

After getting the repeating segment model, each segment of the mixture spectrogram is compared, that is derived in segmentation, with the repeating segment, which is the median of all segments of mixture spectrogram. The element-wise minimum between them is calculated, and if the repeating segment is smaller than a segment of the mixture spectrogram, replacement of segment with the repeating segment is done. The rationale is that if the value of a segment of the mixture spectrogram is bigger than the repeating segment, it denotes that in this segment, it contains more non-repeating information. In order to remove the non-repeating pattern, this segment needs to be replaced with the repeating segment. Otherwise, if the value of a segment is smaller than the repeating segment, it denotes that this segment contains less non-repeating pattern and it is retained. After comparison and replacement, the new spectrogram derived is called repeating spectrogram. After obtaining the repeating spectrogram, it is started to remove the non-repeating part from the mixture spectrogram. The basic idea is to do time-frequency mask. Repeating spectrogram W is divided by mixture spectrogram V to get the time frequency mask M. If some parts of the mixture spectrogram are similar to the repeating spectrogram, the value of W / V will be near 1 and these parts is counted as music background with repeating pattern. Otherwise, the value of W / V will be near 0 and these parts will be counted as non-repeating singing voice foreground. The mask M contains the repeating information of the mixture spectrogram and all values in mask M are in the range from 0 to 1. Then, multiply the mask M with the original mixture spectrogram V. Since the range of all values of M is from 0 to 1, the music part will be reserved after multiplication while the singing voice is removed. That is to say, the result of M * V is the music spectrogram with singing voice removed.

As discussed earlier REPET can be successfully applied for separation of music/voice. In case of songs with full track variations are seen over time. It is proposed to apply REPET to the signal's local window over time and extend the method to longer musical pieces. REPET is independent of particular statistics (e.g., MFCC or chroma features), free from complex frameworks (e.g., pitch-based inference techniques or source/filter modeling), and free from preprocessing (e.g., vocal/non-vocal segmentation or prior training).

5. LPC MODEL

Analysis and synthesis of speech signals is done using LPC technique. It helps to estimate primary speech parameters like formants, spectra and pitch. Figure 5 shows the LPC Vocoder block diagram.

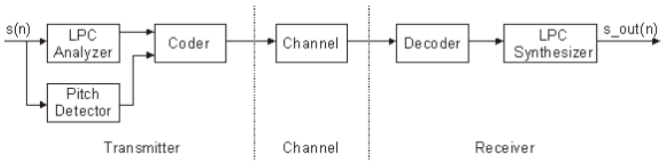

Figure 5: Block diagram of an LPC vocoder.

The sum of squared differences of the original speech signal and the expected speech signal for a finite duration is minimized using LPC and gives predictor coefficients (a_k) unique set. For every frame the coefficients are estimated that are 20ms long normally. Apart from this gain (G) is also important. The time-varying digital filter transfer function is as below:

$$H(z) = \frac{G}{1 - \sum a_{k z^{-k}}}$$

The calculation of the summation is done from the range of k=1 to p, that is 10 for the LPC-10 algorithm, and 18 for the improved algorithm which is utilized. Here the coefficients transmitted to LPC synthesizer are the first 18. The covariance method and the Auto-correlation formulation are the preferred methods of the available for coefficient computation and Auto-correlation formulation is used for implementation as it is better than the covariance method considering the above equations the denominator polynomial is assured to be in the unit circle and confirming system stability H. The desired parameters are calculated using Levinson - Durbin recursion. The simplified speech production model is represented in figure 6.

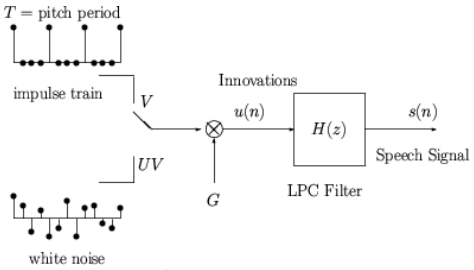

Figure 6 Simplified Speech Production Model

The decision of whether the sound is voiced or unvoiced is a part of LPC analysis of each frame. Impulse train with nonzero taps occurring every pitch period is utilized to represent the sound that is decided to be voiced. Pitch period/frequency is determined by pitch-detection. The pitch period is estimated using autocorrelation function. On the other hand white noise is used with T=0 as the pitch period to represent the sound if decided to be unvoiced. LPC synthesis filter excitation is resulted due to either white noise or impulse train. It is important to note that pitch, gain and coefficient parameters varies with time from one frame to another.

A. LPC speech analysis technique

For an input speech signal the spectral information is generated currently uses LPC speech analysis techniques as most of the voice recognition devices are concerned. Voice recognizers use observation vectors that are created by LPC techniques in the process of voice recognition. To recognize the input speech the comparison of the observation vectors with stored model vectors is carried out. The industries like telephony and consumer electronics uses voice recognition systems. To have Hands Free dialing, or voice dialing in mobile phones voice recognition is used. Windowing techniques are used to analyze the speech signal.

B. Windowing

To have a smooth estimated power spectrum and avoiding the abrupt transitions in frequency response between adjacent frames one may choose to use some kind of windowing function in the linear analysis

Spectral smoothing techniques are used to avoid distinct peaks in the spectrum, which will result in poles near the unit circle.
The multiplication effect of the input with a finite-length window is equal to convolving the power spectrum with the frequency response of the window and caused an averaging effect on the signal power spectrum in the side-17 lobes in the frequency response. The use of 160-sample frames in the linear analysis would be equal to windowing the input with a 160-point rectangular window. In the analysis, a 160-point Hamming window is generally used, which has better frequency properties than the rectangular window. The effect of this is to produce a weighted average of the input, where the 160 samples in the center of the Hamming window correspond to the frame being processed, i.e. the last sub-frame of the preceding frame and the first one of the next frame are also included in the analysis. This alleviates the effect of abrupt transitions in the frequency properties of adjacent frames. There are two important windows which are used frequently for windowing techniques. These are discussed below:

- *Hamming and Hanning Windows*

The window function is expressed as

$$W_2(n) = W_1(n)\left(\beta_1 - 2\beta_2 \cos(\alpha n)\right), \alpha = \frac{2\pi}{L-1}$$

$$W_2\left(e^{jw}\right) = \beta_1 W_1\left(e^{jw}\right) - \left(\beta_2 W_1\left(e^{j(w+\alpha)}\right) - \beta_2 W_1\left(e^{j(w-\alpha)}\right)\right)$$

Where L represents the width, in samples, of a discrete-time and pulling out the exponential terms, and simplifying it will be,

$$e^{j\left(\frac{w(L-1)}{2}\right)}$$

$$S(i,j) = \frac{v_i v_j}{/v_i v_j/}$$

The desired 1 and 2 can be chosen by making the side lobes close to 0. Considering the above sum the first term is out of phase while the second and third term are for every side lobe.

Below are typical choices:

For Hanning window:

First lobe=0.5,

Second lobe=0.25.

For Hamming window: 1= 0.54, 2= 0.23.

The amplitude for the first side lobe is <1% of the main lobe amplitude considering the constraints less than 40dB

- *Hamming Window Characteristics*

Main lobe: 2.4 dB
First null: 2.5 dB
First side lobe: 2.6 dB

6. RESULTS AND DISCUSSION

Here, 40 song clips are considered in the WAVE file form that is sampled at 44.1 kHz and duration from 25 sec to 1 minute for evaluating the REPET. In the process of separation for a mixture the STFT using half-overlapping Hamming windows of N=2048 samples for each window is calculated corresponding to 40 milliseconds at 44.1 kHz.

From the beat spectrum b the repeating period p was estimated automatically by estimating the local maxima in b and identifying the one which periodically repeats most often, with the highest accumulated energy over its periods, and extracts the repeating patterns via time frequency masking. The average computation time for REPET is 0.016 second for 1 second of mixture.

Spectrogram Analysis

Below figures shows the spectrograms for mixture signal, separated background signal (music), and separated foreground signal (voice).

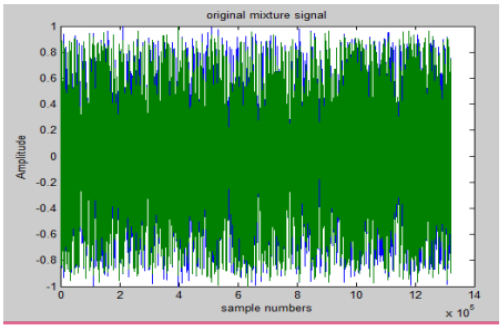

Figure 7: Mixture signal

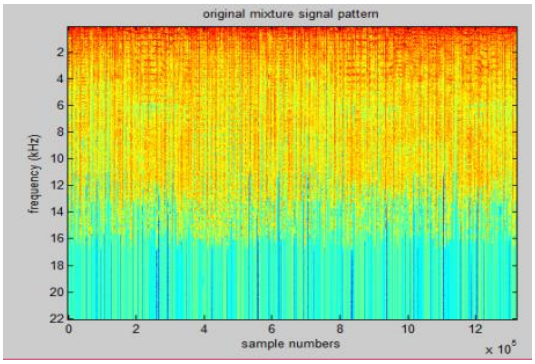

Figure 8: Mixture Spectrogram

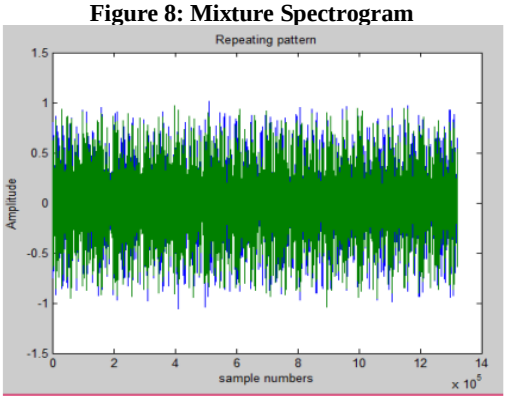

Figure 9: Background (music) signal

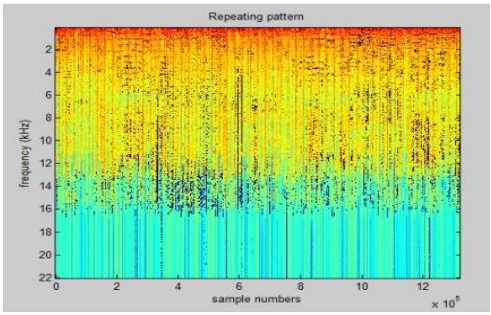

Figure 10: Background (music) Spectrogram

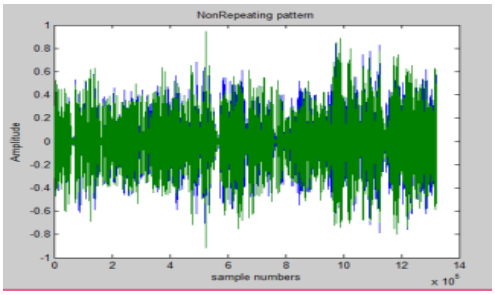

Figure11: Foreground (voice) Signal

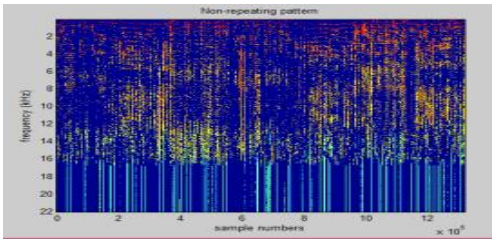

Figure12: Foreground (voice) Spectrogram

From the spectrogram it can be stated that the time-frequency representation is sparse and varied for non-repeating foreground as compared with the repeating background (music) time frequency representation. The background (music) spectrogram is dense and low ranked and foreground (voice) spectrogram is sparse and varied. Some music component in separated voice signal is still assigned in the spectrogram for voice as the parts with high repeating pattern of music are separated.

7. CONCLUSION

In this work, the separation of repeating background from the non-repeating foreground in a mixture is presented using REpeating Pattern Extraction Technique (REPET) which is a simple and novel approach. The idea of the REPET method is to identify the repeating structure in the audio and use it to model a repeating mask. The mask can then be compared to the mixture signal to extract the repeating background. The work tried to propose a method that is novel in the case of music/voice separation by extraction of the underlying musical repeating structure is concerned. A data set of 40 song clips from five different languages is considered. The evaluation of the same presented that the method is able to achieve better performance than the existing automatic approach as far as separation is concerned without the aid of any particular features or complex calculations. The proposed method is simple, fast and completely automatable. By the combination of REPET with the LPC, the average SNR gave better performance to have an improved melody REPET can be used as preprocessor to pitch detection algorithms. For music/voice separation REPET can be efficiently applied. REPET is robust to real-world recordings and can be extended to full-track songs.

8. REFERENCES

[1] Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from background music," in *Proc. Int. Symp. Frontiers of Res. Speech and Music*, Mysore, India, May 8–9, 2007.

[2] 2. Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Oct 2012, pp. 583–588.

[3] 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, January 2013.

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

[5] Palmer and C. L. Krumhansl, "Pitch and temporal contributions to musical phrase perception: Effects of harmony, performance timing, and familiarity," *Perception & Psychophysics*, vol. 41, no. 6, pp. 505–518,1987.

[6] Friberg and S. Ahlba¨ck, "Recognition of the main melody in a polyphonic symbolic score using perceptual knowledge," *J. New Music Research*, vol. 38, no. 2, pp. 155–169, 2009

[7] M. Ramona, G. Richard, and B. David, "Vocal detection in music with support vector machines," in *Proc. Int. Conf. Acoust., Speech, and Signal Process.*, 2008, pp. 1885–1888.

[8] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitchbased methods for background and melody separation," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 22, no. 12, pp. 1884–1893, 2014.