International Journal of Scientific Engineering and Research (IJSER)
www.ijser.in
ISSN (Online): 2347-3878
Volume 2 Issue 3, March 2014

Overview of Performance Enhancement of REpet Algorithm Using MFCC

Snigdha S. Bhattacharjee[1], Amruta Moon[2]

Department of Computer Science and Engineering
G.H.R.I.E.T.W., Rashtrasant Tukdoji, Maharaj, Nagpur University, Nagpur, India

Abstract: ***Repeating Pattern Extraction Technique (REPET)*, a novel and simple approach for separating the repeating "background" from the non-repeating "foreground" in a mixture. The basic idea is to identify periodically repeating patterns in the audio (e.g., a guitar riff or a drum loop), and then separate the repeating "background" from the non-repeating "foreground" (typically the vocal line). This is embodied in an algorithm called *REpeating Pattern Extraction Technique (REPET)*. In this project the performance of the repeating algorithm using the Mel Frequency Cepstral Coefficients (MFCC) with the help of MATLAB will be implemented.**

Keywords: Mel Filter Matrix, Filter Bank, Cepstrum, Spectrum

Abbreviations: MIREX (Music Information Retrieval Evaluation eXchange)

1. Introduction

Repetition "is the basis of music as an art" .The ability is to conveniently isolate a song into its music and voice components are of great interest for a wide range of applications. We take a fundamentally different approach to separate the lead melody from the background accompaniment. The idea is to identify the periodically repeating patterns in the audio (e.g., a guitar riff or a drum loop), and then separate the repeating "background" from the non-repeating "foreground" (typically the vocal line). This is embodied in an algorithm called *REpeating Pattern Extraction Technique (REpet)* using MFCC (Mel Frequency Cepstral Coefficients) [7].

The ability to efficiently separate a song into its music and voice components would be of great interest for a wide range of applications, among others instrument/vocalist identification, pitch/melody extraction, audio post processing, and karaoke gaming. Existing methods in music/voice separation do not explicitly use the analysis of the repeating structure as a basis for separation.

The vital notion is to determine the periodically repeating segments in the audio, compare them to a repeating segment model derived from them, and extract the repeating patterns via time-frequency masking**.** Current trends in audio source separation are based on a *filtering* paradigm, in which the sources are recovered through the direct processing of the mixtures. When considering Time-Frequency (TF) portrayals, this filtering can be estimated as an element wise weighting of the TF representations (e.g. Short-Time Fourier Transform) of the mixtures [6]. When individual TF bins are assigned weights of either 0 (e.g. background) or 1 (e.g. foreground), this is known as binary TF masking.

Cepstral analysis, the historical father of the MFCCs:

Cepstrum is may be the most popular homomorphic processing because it is useful for deconvolution. To understand it, one should remember that in speech processing, the basic human speech production model adopted is a **source-filter** model.

1.1. Source: It is related to the air expelled from the lungs. If the sound is **unvoiced**, like in "s" and "f", the glottis is open and the vocal cords are relaxed. If the sound is **voiced**, "a", "e", for example, the vocal cords vibrate and the frequency of this vibration is related to the pitch.

1.2. Filter: It is responsible for giving a shape to the spectrum of the signal in order to produce divergent sounds. It is related to the vocal tract organs.

1.3.Roughly speaking: A good parametric representation for a speech recognition system tries to eliminate the influence of the source (the system must give the same "answer" for a high pitch female voice and for a low pitch male voice), and symbolize the filter. The ***problem*** is: source e (n) and filter impulse response h (n) are convoluted. Then we need deconvolution in speech recognition applications [2].

2. Literature Review

The justification for this approach is that many musical pieces are composed of structures where a singer overlays varying lyrics on a repeating accompaniment. Examples include singing different verses over the same chord progression or rapping over a repeated drum loop [1]. Zafar Rafii and Bryan Pardo [1] compared this extended REPET with Durrieu's system [2] enhanced with the unvoiced lead estimation. An analysis window of 46.4 milliseconds, an analysis Fourier size of N=2048 samples, a step size of 23.2

milliseconds, and a number of 30 iterations has been used. Also a high-pass filtering of 100 Hz on the voice estimates for both methods, and use the best repeating period for REPET has been applied [1].

If a model of the voice is being available, then TF bins are classified under the music if the corresponding observations are far from the model, thus defining a binary mask. With this in mind, a recently proposed technique called REPET (REpeating Pattern Extraction Technique) focuses on modeling the *accompaniment* instead of the vocals. For longer musical excerpts however, the musical background is likely to vary over time, limiting the length of excerpt that REPET can be applied to. Furthermore, the binary TF masking used in REPET leads to noise artifacts [2].

When dealing with polyphonic signals, two strategies can be undertaken: either the whole signal is processed with a direct extraction of information, or it is split in several individual components ideally hypothesized as monophonic streams. An alternative strategy is emerging, which consists in defining a mid-level representation that facilitates the subsequent processing (for tempo estimation, for instrument recognition and pitch estimation, or for genre classification). Mid-level representations are often viewed as a signal transformation from which a collection of indicators are extracted and indexed by their time instants [3]. Source separation is another very intense field of research where the objective is to recover several unknown signals called *sources* that were concocted in observable *mixtures*. Source separation problems can be classified as sound processing, telecommunications and image processing. If fewer mixtures exist than sources, then the problem is said to be *underdetermined* and is notably known to be very difficult [4].

The objective of voice recognition is to determine which speaker is present based on the individual's utterance. In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency. MFCC is having two types of filters, viz., spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz. A subjective pitch which is present on Mel Frequency Scale is used to capture important characteristic of phonetic in speech [5]. The approach is based on sparse shift-invariant probabilistic latent component analysis (SI-PLCA), a probabilistic variant of convolutive non-negative matrix factorization (NMF) in which the algorithm treats a musical recording as a concatenation of a small subset of short, repeated patterns, and is able to simultaneously estimate both the patterns and their repetitions throughout the song. The analysis naturally identifies the long-term harmonic structure in a song, whereas the short-term structure is encoded within the patterns themselves [6].

The monaural singing voice separation methods can be generally classified into two categories depending on their underlying methodologies: the spectrogram factorization and the pitch-based inference. Pitch-based inference methods, on the other hand, use the extracted vocal pitch contours as a cue to separate the harmonic structures of the singing voice [7].

An important problem in separation of pitched musical sounds is the estimation of time–frequency regions where harmonics overlap. Therefore a sinusoidal modeling-based separation system has been proposed that can effectively resolve overlapping harmonics. The strategy is based on the observations that harmonics of the same source have correlated amplitude envelopes and that the change in phase of a harmonic is related to the instrument's pitch [8].

Klapuri proposed polyphonic signals which are derived from computationally efficient fundamental frequency (F0) estimator [9]. Based on the outcome, three different estimators namely: a "direct" method, an iterative estimation and cancellation method, and a method that estimates multiple F0s jointly are proposed. The number of concurrent sounds is estimated along with their F0s [9].

A thorough review of the main methods and an original categorization based on speed, memory requirements and accuracy has been proposed [10] in which an algorithm is presented for the estimation of the fundamental frequency (F0) of speech or musical sounds is based on the well known autocorrelation method with a number of modifications that combine to prevent errors. It is based on a signal model (periodic signal) that may be extended in several ways to handle various forms of aperiodicity that occur in particular applications [11].

The beat spectrum is calculated from the audio using three principal steps. First, the audio is parameterized using a spectral or other representation. The beat spectrum results from finding periodicities in the similarity matrix, using diagonal sums or autocorrelation [12]. A new system that automatically generates audio thumbnails for selections of popular music. Our system employs a feature-classification framework for audio analysis. This feature class represents the spectrum in terms of pitch-class, and is derived from the chromagram [13].

A set of ten mel-frequency cepstrum coefficients are computed every 6.4 ms resulted in the best performance, viz 96.5 percent and 95.0 percent recognition with each having two speakers. The brilliant performance of the mel-frequency cepstrum coefficients may be attributed to the fact that they better represent the perceptually relevant aspects of the short-term speech spectrum [14].

3. Research Methodology to be Employed

3.1 Proposed Research

We propose an enhanced repeating period estimation algorithm, an improved repeating segment modeling, and an alternate way for building the time-frequency masking. We also propose a simple procedure to extend the method to longer musical pieces. An algorithm to estimate the repeating period had been developed and implemented:

Finding the repeating period from the beat spectrum using MFCC (Mel Frequency Cepstral Coefficients).

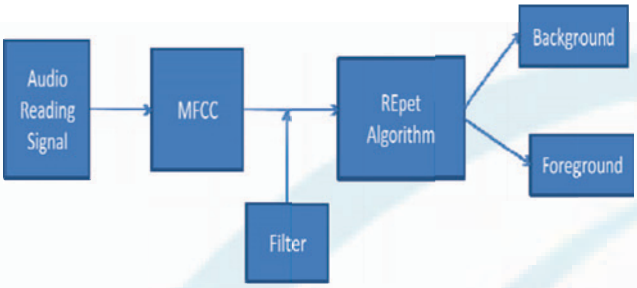

Figure 1: Block Diagram of the Overall Design Analysis of Music & Voice Segregation

3.2. Proposed Methodology

The proposed methodology can run into three modules:

a) Melody Extraction and Music Structure Analysis
b) Music/Voice Separation Using MFCC Components
c) Sound Separation using MFCC and REPET Algorithm

a) Melody Extraction and Music Structure Analysis
REPET can consequently improve melody extraction, by using it to first separate the repeating background, and then applying a pitch detection algorithm on the voice estimate to extract the pitch contour. We employ two different pitch detection algorithms: the well-known single fundamental frequency (F0) estimator YIN proposed by de Cheveigné, and the more recent *multiple* estimator proposed by Klapuri. YIN is an F0 estimator designed for speech and music, based on the autocorrelation method. And a multiple F0 estimator designed for *polyphonic* music signals, based on an iterative estimation and cancellation of the multiple F0s.

Foote introduced the *similarity matrix*, a two-dimensional matrix where each bin measures the (dis)similarity between any two instances of the audio. The similarity matrix (or its dual, the distance matrix) can be built from different features, such as the Mel-Frequency Cepstrum Coefficients (MFCC), the spectrogram, the chromagram, the pitch contour, or other features, as long as similar sounds yield similarity in the feature space. Different similarity (or distance) functions can also be used, such as the dot product, the cosine similarity, the Euclidean distance, or other functions.

Foote *et al.* developed the *beat spectrum*, a measure of acoustic self-similarity as a function of the time lag; by using a similarity matrix built from the spectrogram .Other beat estimation methods include Pikrakis *et al.* who built a similarity matrix using MFCCs.

b) Music/Voice Separation Using MFCC Components

Music/voice separation methods typically first identify the vocal/non vocal segments and then use a variety of techniques to separate the lead vocals from the background accompaniment, including spectrogram factorization, accompaniment model learning, and pitch-based inference techniques. Also MFCC is used to distinguish frequency, pitch, melody, angle, cepstrum and the peak frequency between any two or multiple audio input signals.

c) Sound Separation using MFCC and REPET Algorithm

Here we develop two types of algorithms so as to segregate the music and voice.

3.2.1 MFCC algorithm

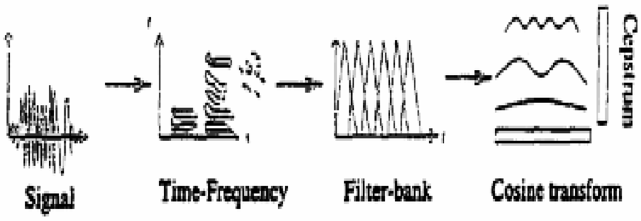

Figure 2: Pictorial representation of Mel-frequency cepstrum (MFCC) calculation

The feature extraction is usually a non-invertible (lossy) transformation, as the MFCC described pictorially in Figure 2 as proposed by Davis [11]. Making an analogy with filter banks, such transformation does not lead to perfect reorganization, which means with only the features it is not possible to reconstruct the original speech used to generate those features. The greater the number of parameters in a model, the greater should be the training sequence.

3.2.2 REPET algorithm

Stage 1: calculation of the beat spectrum b and estimation of the repeating period p. Stage 2: segmentation of the mixture spectrogram V and computation of the repeating segment model S. Stage 3: derivation of the repeating spectrogram model W and building of the soft time-frequency mask M.

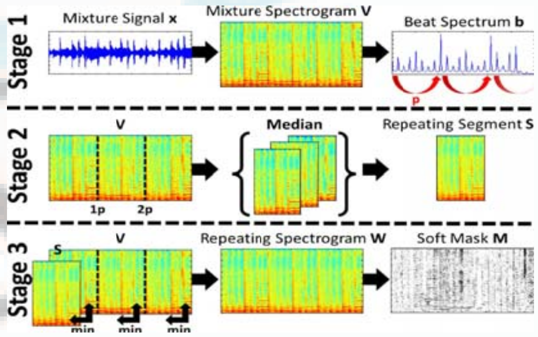

Figure 3: Overview of the REPET algorithm

3.3. Combination of REpet and MFCC to get optimized outputs

The estimation of the parameters of such models is done using tensor factorizations and separation is then consistently performed through the use of an adaptive Wiener-like filter.

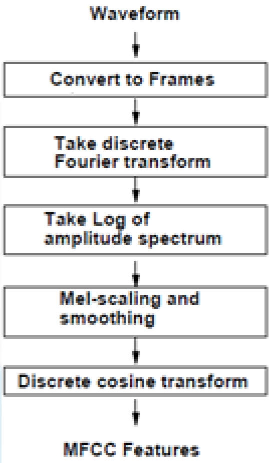
Waveform
Convert to Frames
Take discrete Fourier transform
Take Log of amplitude spectrum
Mel-scaling and smoothing
Discrete cosine transform
MFCC Features

Figure 4: Process to create MFCC Features

The first function is to divide the speech signal into frames, which is done by applying windowing function at fixed intervals. The aim is to model small (typically 20ms) sections of the signal that are statically stationary. The next step is to take the Discreet Fourier Transform (DFT) of each frame. We then retain only the logarithm of the amplitude spectrum because the perceived loudness of the sound has been found to be approximately logarithmic.

The next step is to smooth the spectrum and emphasize a perceptually meaningful frequency which is achieved by collecting the (say) 256 spectral components into (say) 40 frequency bins. Thus the bin frequency follows the so called 'Mel' frequency scale. Spectral components are averaged over Mel-spaced bins to produce a smoothed spectrum. In the speech community the Karhunen-Loeve(KL) or equivalently Principal Components Analysis(PCA) transform is approximated by the Discreet Cosine Transform(DCT) in which cepstral features are obtained in each frame.

Unlike other separation approaches, REPET neither depends on particular statistics, nor requires preprocessing nor does it relies on complex frameworks. Because it is only based on self-similarity, it has the advantage of being elementary, agile, and visionless. It is therefore, completely and easily automatable.

4. Expected Outcome

This specification applies to an input system wherein music and voice can be separated. By using MFCC, segregation can be done in a better way than the REpet Algorithm. Foreground and Background will be produced as outputs. Secondly, parameters like accuracy, SNR, and delay would be found in which each filter output is the sum of its filtered spectral components.

Also by using MFCC comparison of frequencies, pitches, melodies etc can be done between two or multiple input audio signals. Widening the filters in the MFCC filter bank increases recognition for clean speech and provides robust performance in additive white noise.

References

[1] Zafar Rafii,, IEEE, and Bryan Pardo, "Repeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 1, January 2013.

[2] Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard, "Adaptive filtering for music/voice separation exploiting the repeating musical structure," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Kyoto, Japan, Mar. 25–30, 2012, pp. 53–56.

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

[4] Antoine Liutkus, Roland Badeau, and Gäel Richard, "Gaussian Processes for Underdetermined Source Separation", IEEE Transactions On Signal Processing, Vol. 59, No. 7, July 2011.

[5] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal Of Computing, Volume 2, Issue 3, March 2010, ISSN 2151-9617.

[6] R. J. Weiss and J. P. Bello, "Identifying repeated patterns in music using sparse convolutive non-negative matrix factorization," in Proc. 11th Int. Soc. Music Inf. Retrieval, Utrecht, The Netherlands, Aug. 9–13, 2010.

[7] C.-L. Hsu and J.-S. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," IEEE Trans. Audio, Speech, Lang. Process., vol. 18,no.2, pp. 310–319,Feb. 2010.

[8] Yipeng Li, John Woodruff, and DeLiang Wang, "Monaural Musical Sound Separation Based on Pitch and Common Amplitude Modulation", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 17, No. 7, September 2009.

[9] Klapuri, "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes," in Proc. 7th Int. Conf. Music Inf. Retrieval, Victoria, BC, Canada, Oct. 8-12, 2006, pp. 216-221.

[10] Massimo Piccardi, "Background subtraction techniques: a review", IEEE International Conference on Systems, Man and Cybernetics, 2004.

[11] de Cheveigne, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., Vol. 111, No. 4, pp. 1917-1930, April 2002.

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

[13] M. A. Bartsch, "To catch a chorus: using chroma-based representations for audio thumbnailing," in Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust., New Paltz, NY, Oct. 21-24, 2001 , pp.15-18.

[14] Davis, S. and Merlmestein, P., "Comparison of Parametric Representations for Monosyllabic Word

Recognition in Continuously Spoken Sentences", IEEE Trans. on ASSP, Aug, 1980. pp. 357-366.

[15] Beth Logan, “Mel Frequency Cepstral Coefficients For Music Modeling”, Cambridge Research Laboratory, Compaq Computer Corporation, One Cambridge Center,Cambridge,pp.1-11.

[16] Sunil Kumar Kopparapu, Meghna A Pandharipande, G Sita, “Music and Vocal Separation Using Multiband Modulation Based Features”, in IEEE Symposium on Industrial Electronics and Applications (ISIEA 2010), Penang, Malaysia, October 3-5, 2010, pp. 733-737.