A Hybrid Method for Music & Voice Separation - A Periodic Pattern Extraction Technique for Audio Signal

Pooja Gautam[1] B S Kaushik[2]
[1]Student [2]Manager
[1]Department of Electronics & Telecommunication Engineering [2]Department of Electrical & Instrumentation Engineering
[1]RCET, Bhilai (C.G.) India [2]LafargeHolcim India, Bhilai (C.G.) India

Abstract— The separation of music & voice from a song is always being an interesting subject. Various methods for the separation of voice & music have been proposed by the researchers in past few decades. Among them three most popular methods of separation are based on "repeating pattern extraction", "Model-based" & "Pitch based" Although all methods give good results but some of their individual draw back led us to discovery of new methods for separation. The present work started with the review of the existing techniques. Some advantages and disadvantages in their features and performance were pointed out. From the review two methods i.e "Repeating pattern extraction method & Pitch based method" are concluded for the further study. In this view a hybrid approach scheme is proposed for music and voice separation which is not often found in literature. Hybrid approach of REPET & Pitch based method for music/voice separation systems can be used to improve separation performance.

Key words: Music & Voice Separation, A Periodic Pattern Extraction Technique, Audio Signal

I. INTRODUCTION

The concept of music usually begins with an idea that music is a organized sound. A concise definition of music is fundamental to being able to categorize, discuss and consider the phenomenon we understand as being music, which is a key question in the philosophy of music i.e. This characterization is too broad, since there are many examples of organized sound which are not music, such as human speech, and the sounds made by non-human animals and machines.

The voice consists of sound made by human being using the vocal fold for talking, singing, laughing, screaming etc. In human sound production the human voice is specifically a part,in which the vocal folds (or vocal cords) are the primary sound source.

The problem of trying to separate vocals from instrumentals in a song is therefore referred to Music/voice separation so to produce an acappella track which contains only vocals and instrumental/musical track containing only instrumentals sound. For researcher's some of the application includes:

Studying MIR (Music Information Retrieval): It could also be used in Active Noise Control (ANC) for removing periodic interferences, Applications includes: Cancelling periodic interferences in electrocardiography (e.g., the power-line interference) & In speech signals (e.g., in an aircraft a pilot communicating by radio.)

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 remainder of this paper followed as literature survey is given in section II, section III gives conclusion of literature review. Section IV gives an idea about problem related to voice & music separation. Section V different methods related to separation is discussed in this section. Section VI gives the result of various methodologies which are reviewed and section VII concludes the paper.

II. LITERATURE REVIEW

A. Hsu et al. (2012)

Proposed a pitch based separation system. A trend estimation algorithm, estimating the pitch range of the singing voice then the estimated trend is incorporated in the tandem algorithm for acquiring the initial estimate of the singing pitch. According to the initially estimated pitch singing voice is then separated The above two stages, i.e., pitch determination and separation of voice then performs iteration until convergence. A post processing stage is introduced. i.e., which actually decides the pitch contours belonging 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.

Further the upper pitch boundary of singing is as high as 1400 Hz for soprano singers while the pitch bounds of normal speech is between 80 and 500 Hz. The differences of pitch make the separation potentially more challenging.

B. Ozerov et al. (2007)

Proposed a model based method, Models of the sources match precisely the statistical properties of the combined signals. However, it is not possible to train such models always. To overcome this problem, an adaptation scheme was resorted for adjusting the source models in accordance with actual properties of the signals which is observed in the mix.

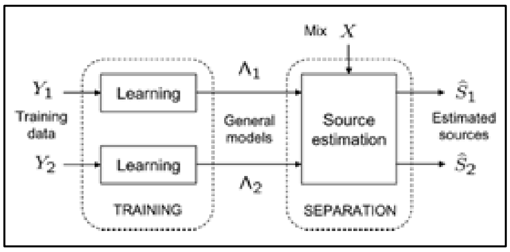

Fig. 1: Source separation depending on general a priori probabilistic models

Representing each source by a GMM is the idea behind these techniques which is composed by a set of characteristic spectral patterns. Each GMM has learning on a training set, which contains sample piece of the corresponding audio class (for instance, speech, music, drums, etc.).

Fig. 1: Represents the principles of probabilistic source separation in general. The general models A1and A2 are trained independently on sets of examples Y1 and Y2. The source estimates S1^ and S2^ are obtained by filtering the mix X with masks estimated from the general source models A1 and A2 the mix itself X.

First, the processed song must contain non vocal parts of reasonable length in order to have enough data for the acoustic adaptation of rhythmic/ music model. Further, the instrumental from non vocal parts should be quite similar to that from vocal parts. Finally, it is preferred to be only single singer at a time, i.e., no chorus or back vocals. At first sight, a majority of well known songs verify these assumptions.

C. Smaragdis and Brown (2003)

Present a methodology for polyphonic music transcription system for modelling, analyzing and Separation of polyphonic musical passages. A harmonically fixed spectral profile (like piano notes) is exibited by any musical instrument. Taking this advantage, the paper models the audio content of the musical passage by a linear basis transforms and use non-negative matrix decomposition method. Music passages from instruments with notes is required that exhibit a static harmonic profile.

D. RAFI and Pardo (2012)

Proposed a method on the assumption that Repetition can be considered as the fundamental element in generating and perceiving structure in music. This method separates the musical background from foreground in a mixture, Instead of looking for periodicities; the method used a similarity matrix for identifying the repeating. Calculation of a repeating spectrogram model is done using the median and further extracting the repeating patterns using a time-frequency masking.

Proposed system doesn’t support the small melodious patterns, but rhythmic patterns have importance for the balance of the music, and can be a way to identify a song.

E. Huang et al. (2012)

Proposed a method that music components can be assumed to be in a low-rank subspace, due to its repetition structure; on the other hand, singing voice are relatively sparse within songs, based on this assumption use Robust Principal Component Analysis, a matrix factorizing method for solving underlying low-rank and sparse matrices.

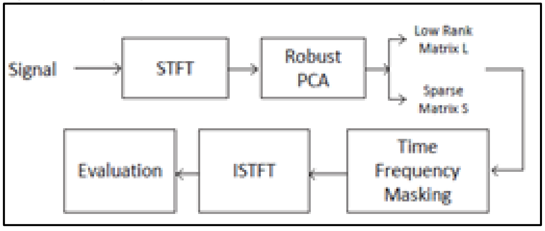

Fig. 2: RPCA Framework

RPCA is a convex program, it is for recovery of low-rank matrices when a fraction of their entries, corrupted by errors, i.e., when the matrix is sufficiently sparse.

Method performs the separation as follows: First computing the spectrogram of music signals as say matrix M, estimated from the Short-Time-Fourier Transform (STFT). Second using inexact Augmented Lagrange Multiplier (ALM) method, which is an efficient algorithm for solving Robust Principal Component Analysis problem, to solve L + S = |M|, given input magnitude of M. Then by RPCA, we can thus obtain two output matrices L and S.

First, if the matrix S is sparser, there is less interference in the matrix S; however, deletions of the actual signal components might result in artifacts. On the contrarily, if S matrix is less sparse, the signal contains fewer artifacts, but from the other sources there is more interference that exist in matrix S.

Secondly, Higher gain factor thus results in lower power sparse matrix S. Therefore, there is larger interference and lesser artifacts at high gain and vice versa.

F. 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 determined. Then the spectrogram is being segmented at period boundaries averaging of segments done to create a repeating segment model. Conclusively, comparison of each and every time frequency bin with the model is done, and the mixture is partitioned using binary time-frequency masking.

Cases where repetitions also happen without a fixed period.

G. RAFI and Pardo (2013)

Proposed new method unlike above previous approaches; this method does not depend on particular features, and also not rely on complex frameworks or calculations and does not also require prior training. Because it is only based on self-similarity, this method can potentially be work on any audio, as long as there is a repetitious structure in the mixture. It has therefore the advantage of being simple, fast, blind, and 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.

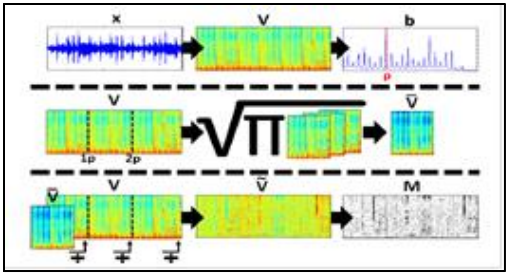

Fig. 3: REPET clearly states the procedure

H. Rafii et al. (2014)

By combination of two unique methods, method was proposed based on Repet & Pitch based methods:

In Parallel combination, from given a mixture spectrogram, REPET derives a back ground mask and 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 acquired 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 mask also known as left over mask. The final background mask & the final melody mask are then derived by weighting and Wiener filtering (WF) the masks.

III. Conclusion of Literature Review

Number of methods applied for separating the repeating harmonics from the non-repeating vocal parts in a composite signal for monaural singing voice separation, existing approaches can be generally classified into three depending on their following methodologies: spectrogram factorization method, model-based methods, and pitch-based methods.

A. Spectrogram Factorization

In this method of music & voice separation the music accompaniment can be supposed to be in a low-rank subspace, on the other hand, singing voices can be considered as relatively sparse within songs, also the repeating property of music has been utilised to separate the music & voice based on this assumption that different methods like RPCA /REPET is used to solve underlying low-rank and sparse matrices.

B. Model-based methods

In this method the property of any lyrical instrument is utilised, any musical instruments exhibit a harmonically fixed spectral profile. Taking advantage of this unique note structure, the model of the audio content of the musical passage is prepared by a linear basis transform and use distinct methods like Adaptation of Bayesian Models / Non-Negative Matrix Factorization is used to extract those music models from mixture

C. Pich-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 two stages above, pitch determination and voice separation then iterates until convergence.

D. Hybrid Model

Till now various papers are presented, work is being done for the separation of voice & music, but the method of Zafar RAFI and Bryan Pardo on the REPET gives the best result, same can be found in literature, the only drawback of this work is that it fails to separate the non repeating beats and the non repeating beats of musical instruments as it is lying in to voice signal. On the other hand the Chao-Ling Hsu and DeLiang Wang Fellow method of pitch based is best suited for the separation of higher pitch value, so above two methods, pitch based & REPETS is selected for Hybrid model.

IV. Problem Definition

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.

Model-based method is suitable for extract the repeating pattern but is requires training.

REPET method also discriminates the repeating pattern and gives the higher values of SDR & GNSDR compared to all other known. The REpeating Pattern Extraction Technique separates the repeating audio signal from the non-repeating audio signal in a mixture. The basic thought is to recognize the periodically repeating segments in the audio, comparing them to a repeating segment model derived from them, and then extracting 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.

V. Methodologies

A hybrid method for Voice & Music separation based on REPET and Pitch based is being used. In original music first we apply pitch method to find out the foreground & background fundamental frequency pitch contors F0 's then segmented REPET method on the F0's of foreground & background.

A. Pitch Based Method

Studies on humans can focus on the melody in musical mixtures by attending to the pitch structure of the audio. From these findings, we have choose to extract the melody by using a pitch-based method that can derive a harmonic mask from identification of the predominant pitch contour in the mixture. Assuming that the melody is the predominant harmonic component in the mixture, pitch-based methods typically first determine the predominant pitch contour by using a pitch detection algorithm, and then infer the corresponding harmonics by estimating the integer multiples of the predominant pitch contour. In this work, we chose a pitch-based method that will be referred to as Pitch. Pitch uses a multi-pitch estimation approach [9] to identify the pitch contour of the singing voice. Although the method originally proposed for multi-pitch estimation of general harmonic mixtures, the algorithm systematically evaluates for predominant pitch estimation and shown to work well compared with other melody extraction method [7]. In this work, we modified the method in [9] to better suit it for melody extraction. While other best approaches to melody extraction there exist (e.g., Hsu et al. [7]), the focus of this work is on combining a simple and clear pitch-based method with a simple and clear rhythm-based method, rather than a comparison of pitch-based methods for source separation. Therefore, we selected a known-good method for which we have a deep understanding of the inner workings and access to the source code. The method can be summarized as follows. First, it identifies peaks in every spectrum of the magnitude spectrogram of the mixture using the method in [8], also defining non-peak regions, and estimates the predominant, from the peaks and non-peak regions. Then, it forms pitch contours by connecting pitches that are close in time (in adjacent frames) and frequency (difference less than 0.3 semitone). Small time gaps (less than 100 milliseconds) between two successive pitch contours are filled with their average pitch value so that the two contours are merged into a longer one, if their pitch difference is small (less than 0.3 semitone). Shorter pitch contours (less than 100 milliseconds) are removed. This is to remove some musical noise caused by pitch detection errors in individual frames. Since some estimated pitches may actually correspond to the vocal instead of the melody, we used a simple method to discriminate pitch contours of melody and accompaniment, assuming that melody pitches vary more (due to vibratos) than accompaniment pitches [60]. More specifically, we calculated the pitch variance for each pitch contour, and removed the ones whose variance is less than 0.05 square semitones. The remaining pitch contours are supposed to be

B. Repeating Pattern Extraction Technique (REPET)

Repetition in each music structure is its basic principle. Any musical pieces being characterized by an underlying repetitive structure over which varying elements are superimposed.

The basic idea is to:

- Identify the periodically repeating segments,
- Repeating segment modeling, and
- Extract the repeating patterns via time-frequency masking.

C. Identify the Periodically Repeating Segments

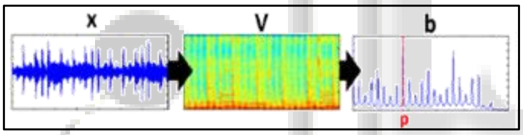

Fig. 4: Identify the Periodically Repeating Segments

Periodicities in any mixture signal can be found by using the autocorrelation, measuring the similarities between segments and lagged version of itself over the successive intervals of time.

Given a mixture signal x, method first calculate its Short-Time Fourier Transform (STFT) X, by using half-overlapping Hamming windows of N samples. Then derives magnitude spectrogram V by taking absolute values of the elements of X, while keeping the DC component and discarding the symmetric part of segment. Then computing autocorrelation of each row of the power spectrogram V2 (element-wise square of V) and obtain the matrix B. Method use V2 to emphasize the appearance of peaks of periodicity in B. If the mixture signal x is stereo, then averaging of V2 over the channels. 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).

D. Repeating Segment Model

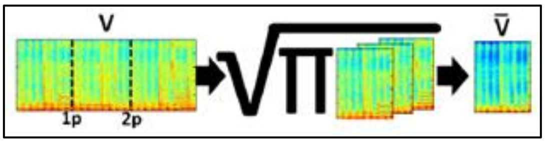

Fig. 5: Repeating Segment Model

After estimation the period p of the repeating musical structure, the method uses it to evenly segment the spectrogram V into segments of length p. Then computing mean repeating segment V over r portion of V, which can be thought of as the repeating segment model. The approach is that time-frequency bins comprises the repeating patterns had similar values at each period that would also be similar to the repeating segment model. Experiments had shown that the geometric mean lead to a effective extraction of the repeating musical structure than arithmetic mean.

E. Binary Time-Frequency Masking

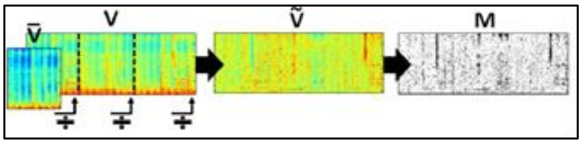

Fig. 6:Binary Time-Frequency Masking

After computing the mean repeating segment V, method divides each time-frequency bin in each segment of spectrogram V by the corresponding bin in V . Then taking the absolute value of the logarithm of each bin to get a modified spectrogram ~V and furthermore the repeating musical structure generally involving variations. Therefore, method introduce a tolerance t when creating the binary time frequency mask M. Experiments shows tolerance of t = 1 giving good separation results, both for music and voice.

Once the binary time-frequency mask M is computed, then symmetrising and applying to STFT X of the mixture signal x to have the STFT of the music and the STFT of the speech. The music signal and voice are finally achieved by inverting their corresponding STFTs into the time domain.

F. Hybrid Method

A hybrid method for Voice & Music separation based on REPET and Pitch based method will be used, the Flow Diagram of hybrid approach is shown below, from the mixture signal spectrogram we will first find out the repeating segment & period of repetition, same as REPET method and separates the Voice & music part by time frequency masking.

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.

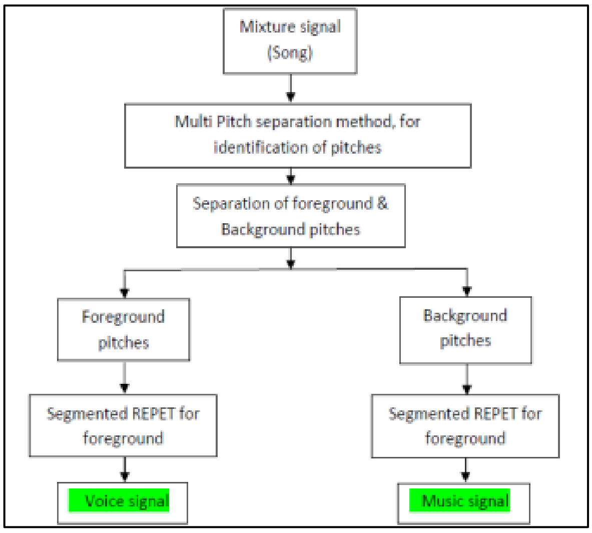
Mixture signal (Song)
Multi Pitch separation method, for identification of pitches
Separation of foreground & Background pitches
Foreground pitches
Background pitches
Segmented REPET for foreground
Segmented REPET for foreground
Voice signal
Music signal

Fig. 7: Hybrid Approach flow diagram.

VI. PERFORMANCE EVALUATION

In this section, we judge REPET on a data set of 100 song clips, compared with a recent competitive singing voice separation methods. We first introduce the data set and the competitive method. We then present the performance measures and the comparative results:

A. Data Set

Hsu et al. proposed a data set called MIR-1K1. The data set consists of 1,000 song clips in the form of split stereo WAVE files sampled at 16 kHz, extracted from 110 karaoke Chinese pop songs, performed mostly by amateurs, with the music and voice recorded separately on the left and right channels, respectively. The duration of the clips ranges from 4 to 13 seconds. The data set also holds manual annotations of the pitch contours, indices of the vocal/non-vocal frames, indices and types of the unvoiced vocal frames, and lyrics. Following the framework adopted by Hsu et al. in [7], we used the 1,00 song clips of the MIR-1K data set.

B. Competitive Method

Repeating Pattern Extraction Technique (REPET):A Method for Music/Voice Separation by Raffi is being chooses as competitive method for performance comparison.
http://sites.google.com/site/unvoicedsoundseparation/mir-1k
http://www.music-ir.org/mirex/wiki/MIREX_HOME
http://bass-db.gforge.inria.fr/bss_eval/

C. Performance Measures

To measure performance in source separation, Févotte et al.designed the BSS_EVAL toolbox3. The toolbox proposes a formula for SDR that was intended to quantify the quality of the separation between a source and its estimate.

VII. RESULT

The performance of various methods can be evaluated by Measuring the separation quality between the estimated voice and the original voice by Signal-to-Distortion Ratio (SDR)

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

Higher values of SDR mean better separation.

Methods	Foreground			
	Mens Voice	Womens Voice	Mens Voice with Noise	Womens Voice with Noise
	SDR			
REPET	12.46	18.88	13.82	16.43
Hybrid Model	17.35	25.89	18.24	20.97

Table 1: A Comparison of Performances Foreground

Methods	Background			
	Mens Voice	Womens Voice	Mens Voice with Noise	Womens Voice with Noise
	SDR			
REPET	15.03	8.30	12.16	2.44
Hybrid Model	17.58	11.82	19.65	13.92

Table 2: Comparison of Performances Background

VIII. CONCLUSIONS

We have proposed a novel method for music/voice separation, by extraction of the underlying musical repeating structure. Evaluation on a dataset of 100 song clips showed that this method can achieve better performance in separation than an existing automatic approach, without requiring any particular features or complex calculations. This proposed method also has an advantage of being simple, fast and completely automatable.

The SDR achieved from REPET and Hybrid method is given by table 1 for male and female singer also with noise in mixture and the avg SDR by using hybrid method:

1) For Male singer SDR is found to be 17.35 for foreground & 17.58 for background for voice to music ratio 0 DB, which is higher than REPET method which is 12.46 & 15.03 respectively.
2) For Female singer SDR is found to be 25.89 for foreground & 11.82 for background for voice to music ratio 0 DB, which is higher than REPET method which is 18.88 & 8.30 respectively.
3) For Mixture (song) with noise of male singer, SDR is found to be 18.24 for foreground & 19.65 for background for voice to music ratio 0 DB, which is higher than REPET method which is 13.82 & 12.16 respectively.
4) For Mixture (song) with noise of female singer, SDR is found to be 20.97 for foreground & 13.92 for background for voice to music ratio 0 DB, which is higher than REPET method which is 16.43 & 2.44 respectively.

Hybrid compression gives better performance than REPET (best known method) method. The hybrid method is combination of Pitch & REPET.

REFERENCES

[1] Po-Sen Huang, Scott Deeann Chen, "Singing-voice separation from monaural recordings using robust principal component analysis," Paris Smaragdis, Mark Hasegawa-Johnson IEEE. ICASSP, 2012, pp. 57-60.
[2] Alexey Ozerov, Pierrick Philippe, Frédéric Bimbot, and Rémi Gribonval , "Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs," IEEE transactions on audio, speech, and language processing, vol. 15, no. 5, july 2007, pp. 1564-1578.
[3] Hsu Chao-Ling, Wang D., Roger J. Jyh-Shing, and Hu K , " A Tandem Algorithm for Singing Pitch Extraction and Voice Separation from Music," IEEE transactions on audio, speech, and language processing, vol. 20, no.5, 2012, pp. 1482-1491.
[4] Zafar RAFII, Bryan Pardo, "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, 2011, pp. 1-4.
[5] Zafar RAFII, Bryan Pardo, " Music/voice separation using the similarity matrix," 13th ISIMR, 2012, pp.583-588.
[6] Zafar Rafii, Student Member, IEEE, and Bryan Pardo, Member, IEEE, "REpeating Pattern Extraction

Technique (REPET): A Simple Method for Music/Voice Separation, " IEEE transactions on audio, speech, and language processing, vol. 21, no.1, 2013, pp. 71-82.

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

[8] Zafar Rafii, Zhiyao Duan and Bryan' "Combining Rhythm-Based and Pitch-Based Methods for Background and Melody Separation," IEEE transactions on audio, speech, and language processing, vol. 22, 2014, pp.1884-1893.

[9] Zhiyao Duan, Jinyu Han and Bryan Pardo, "Multi-pitch Streaming of Harmonic Sound Mixtures," Manuscript for IEEE trans. Audio, speech and language processing, 2013, pp.1-13