# A New Approach to Music/Voice Separation Using Resonance-Based Signal Decomposition

**Roozbeh Soleymani, New York University, Polytechnic School of Engineering**


## Abstract

Separation of the voice from background music has been the subject of many recent research studies. As voice and music are mixed in both time and frequency domains, conventional filtering and signal separation methods are not applicable to this problem.

State of the art solutions to music/voice separation take advantage of complex techniques such as pitch detection,texture detection,repeating pattern,temporal and spectral continuity,trend estimation,sparsity and wavelet and voice recognition.However many of the techniques operate based on assumptions and simplifications that are not always valid, thus resulting in high separation error rates.

This poster presents a new approach to music/voice separation based on the use of tunable Q wavelet transform and sparsity_enabled method for resonance-based signal decomposition.Resonant-based signal decomposition method performs well when applied to the separation of high resonant and low resonant components of the audio signals.

**Stage 1**

The fundamental difference between music and voice in this case is that music has a more consistent behaviour when it comes to high and low resonant components and it mostly contains either high or low resonant components rather than a mixture of two(Since there is no exact definition for “Low” or “High” in this case if we carefully tune the tunable Q wavelet transforms we can have most of the music energy as high resonant component most of the time).

As opposed to music ,the human voice does not have such a consistency and most of the time contains a mixture of high and low resonant components at the same time, In this method , this feature has been used to recognize the sections which are pure music from the sections which are the mixture of voice and music by defining a cost function.So we can simply classify the segments with a high percentage of high resonant component as “Music Only” or “Pure Music” parts and the segments which have a considerable amount of low resonant content(compared to previous case) as “Mixture of Voice/music” segments.

**Stage 2**

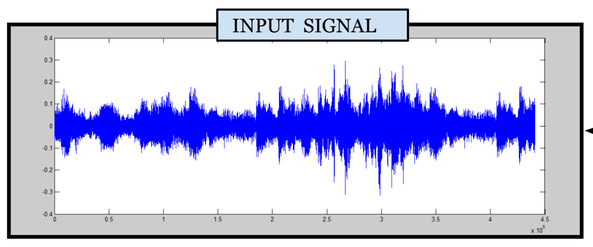
The second stage of this method attemps to restore the rest of the music which is mixed with singing voice . The key information which is very useful in this stage is the music features that are exracted from the segments that have been classified as pure music in the first stage.Both pure music and mixture part are divided into several narrow frequency bands in frequency domain and several short time blocks in time domain to make a profile which contains the infrmation of the signal energy variation in each of these frequency bands over time also frequency arrangement in each time block.At this stage three independent classification methods are applied to each frequency subband to decide weather the subband in mixed section bear any resemblance to its counterpart in pure music section or not.If there is a considerable similarity we can assume than band as a music band as well.The three classification methods are based on frequency accompaniment pattern detection (An extended concept of harmonic detection) ,symmetricity based cross correlation similarity finder, and overlap based similarity finder.The final decision is made based on majority voting from three classifiers.

**Stage 3**

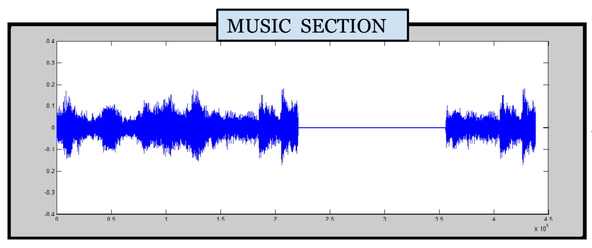
The output of the second stage will separate music and voice signal in mixture part of the input. However both music and voice include low energy remnants of the each other. Even though these components have a very low energy compared to the main signal, they still can be heard by human ear and therefore have a high negative impact on separation quality.

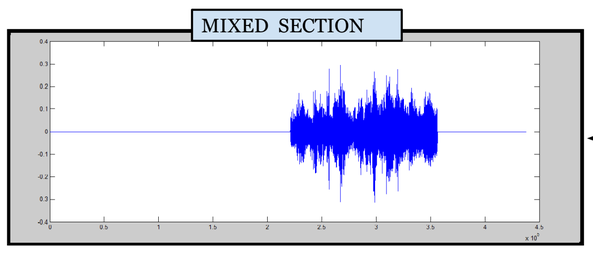
In order to remove these parasitic low energy components in third stage again using the information from the music segments(stage 1) and results of similarity finder methods(stage 2) the time blocks in the music signal which is separated from the voice(and contaminated with some voice remnants) will be substituted with a similar clean pure music time blocks obtained from pure music section.Here, it is assumed that every note which is played in the mixture section has been played at least once in the music section. Which seems to be a relatively reasonable assumption.(Normally singer does not start singing immediately after the music starts).

This approach has been tested with several music/voice signals and whenever the input signal met assumption requirements the results demonstrated almost a perfect separation efficiency. How ever in cases which the input signal does not meet the above mentioned assumption the result is still very good.

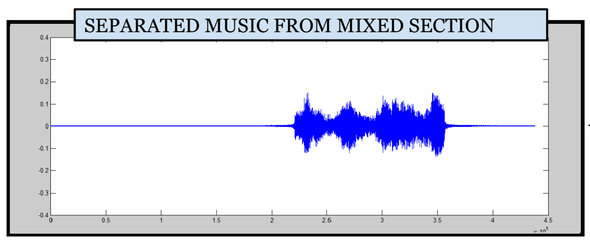



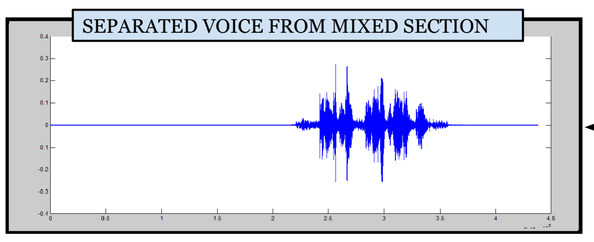

**INPUT SIGNAL**
Mixture of Voice and Music



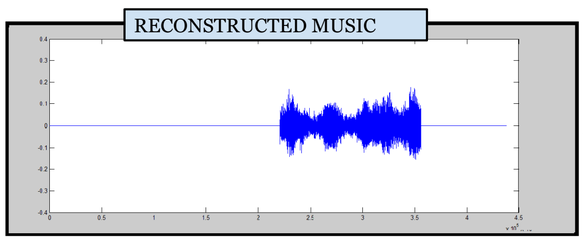



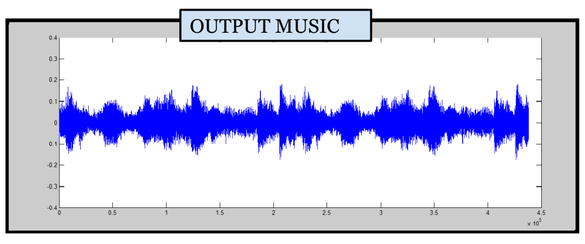

**STAGE 1**
Music/mixture separation using tunable Q wavelet transform



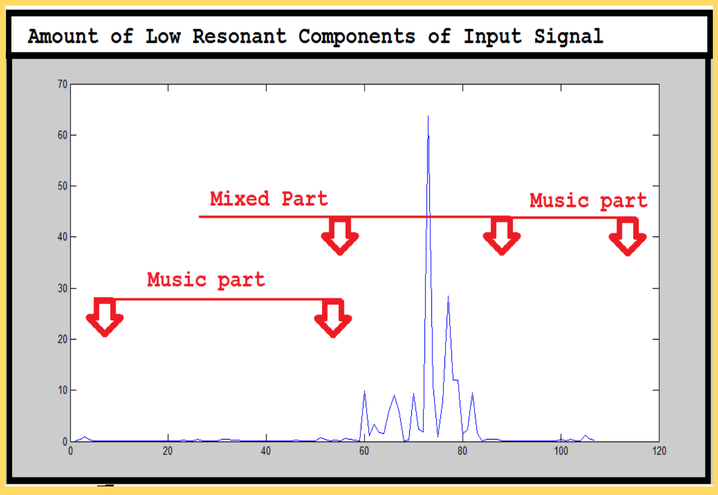

**STAGE 2:**
Separation of music from voice in mixture part using data extracted from pure music



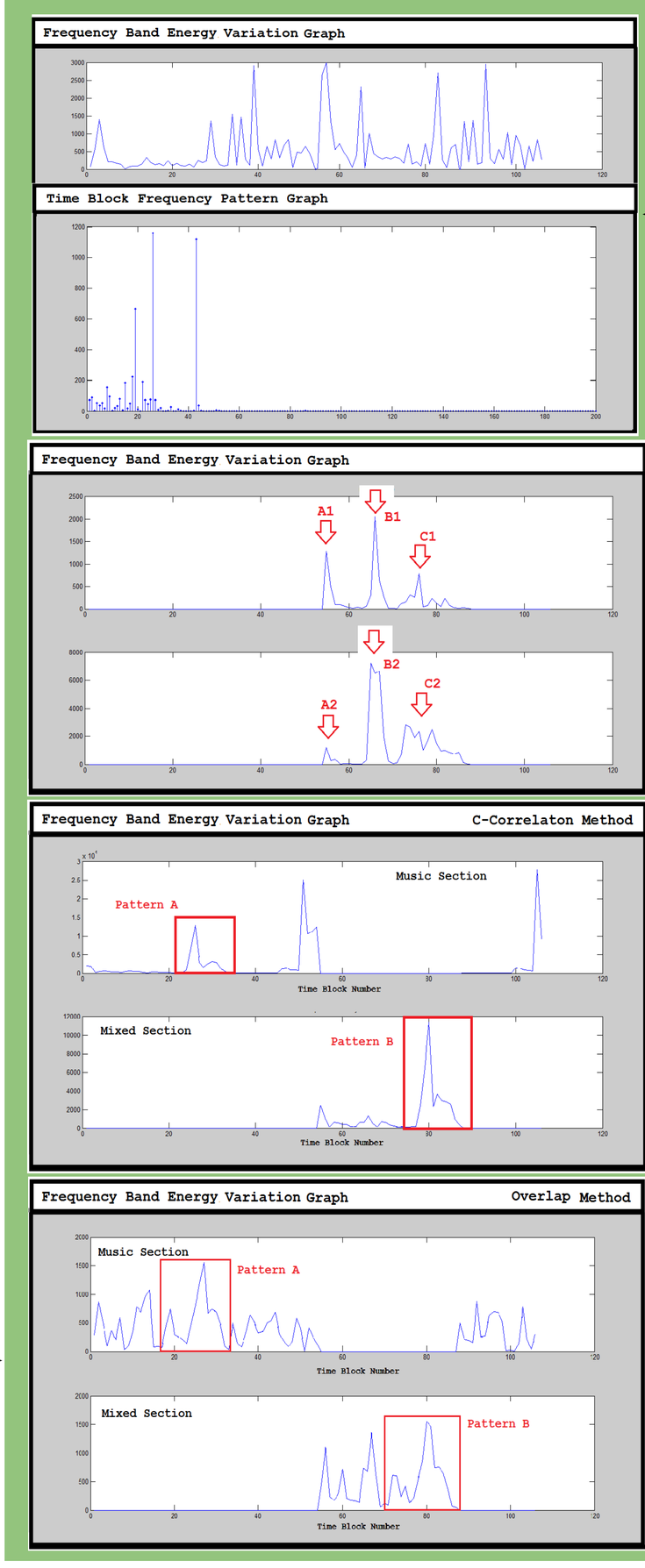

**STAGE 3:**
Substitution of separated music blocks extracted in stage 2 with similar clean music blocks using data in the pure music section

## First Stage




- Selection Of The Parameters For Tunable Q Factor Wavelet Transform
- Splitting The Signal Into Several Short Time Blocks
- Calculating High And Low Resonant Contents Of Each Time Block
- Categorizing Each Block As Pure Music Or Mixture Of Music And Voice Based On Its High/low Resonant Contents
- Connecting The Mixture(voice+music) Time Blocks To Make A Signal Including Mixture Sections Of The Input Signal And Connecting The Pure Music Time Blocks To Make A Signal Including Pure Music Sections Of The Input Signal

## Second Stage




- Split Music Section Into Many Time And Frequency Blocks And Build The Time /frequency Profile
- Split Mixture Section Into Many Time And Frequency Blocks And Build The Time /frequency Profile
- Categorize All Frequency Bands Into High , Medium, Low And Negligible Energy Bands In Each Profile
- Find The Frequency Bands Which Are Categorized As Band With Negligible In Music Sections And Are Medium Or High Energy Level In Voice Part And Categorize Them As Voice Freq Bands
- Find The Time/energy Ratio Pattern Of Accompanying Freq Band For Each Freq Band In Both Profiles
- Find The Bands Which Have Similar Freq. Accompaniment Patterns In Both Music And Mixture Section And Categorize Them As Music Band
- Apply the Remaining Bands to Symmetricity Based Cross Correlation Similarity Finder Method. And Categorize Signal With Similarity Higher Than A Certain Level As Music
- Apply The Remaining Bands to Overlapping Based Similarity Finder Method.categorize Signal With Similarity Higher Than A Certain Level As Music
- Take The Majority Voting Between The Three Above Mentioned Methods to Decide Music Bands
- If There Is Considerable Amount of Energy In Remaining Bands Split The The Remaining Bands Into Narrower Freq Bands And Shorter Time Blocks And Repeat The Above Steps
- Apply The Multi Band Pass And Multi Band Stop Filters To Mixture Signal And Separate Music And Voice Components of The Mixture Signal

## Third Stage

| Mixed Section Frequency Bands | | | | | | |
|---|---|---|---|---|---|---|
| Band Number | | 28 | 29 | 30 | | |
| Music Band 1 | | 8 | 9 | 10 | | |
| Votes | | 3 | 3 | 2 | | |
| Music Band 2 | . . | 53 | 54 | 55 | . . | |
| Votes | | 4 | 5 | 4 | | |
| Music Band 3 | | — | 70 | 103 | | |
| Votes | | — | 8 | 1 | | |
| Other Votes | | — | 2 | — | | |
| Replace Order | | 53 | 54 | 55 | | |

- Take The Frequency Blocks Which Have The Majority Voting Of Similarity Methods
- Determine The Time Blocks Which Two Bands Are Similar In Them
- Make A Profile Of The Time Blocks In Mixed Section And Substitution Alternatives In Pure Music One
- Select The Best Substitution Alternative Based On The Majority Voting And Neighbouring Bands
- Replace The Best Clean Music Blocks With The Dirty Music Blocks
- Use Information From Reconstructed Version Of Music To Enhance The Quality Of The Voice Part


**References**

[1] Ivan W. Selesnick, Wavelet Transform With Tunable Q-Factor. IEEE Transactions on Signal Processing 59(8): 3560-3575 (2011)

[2] Ivan W. Selesnick, Resonance-based signal decomposition: A new sparsity-enabled signal analysis method. Signal Processing 91(12): 2793-2809 (2011)

[3] David Havelock, Sonoko Kuwano, Michael Vorländer. Handbook of Signal Processing in Acoustics. Springer. pp. 399–417. ISBN 978-0-387-77698-9.

[4] Zafar Rafii, Bryan Pardo, Repeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation. IEEE Trans. on Audio, Speech & Language Processing 21(1): 71-82 (2013)

[5] Chao-Ling Hsu and Jyh-Shing Roger Jang. 2010. On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset. Trans. Audio, Speech and Lang. Proc. 18, 2 (February 2010), 310-319