ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 2, April 2014

An Efficient Extraction of Vocal Portion from Music Accompaniment Using Trend Estimation

Aisvarya V[1], Suganthy M[2]

PG Student [Comm. Systems], Dept. of ECE, Sree Sastha Institute of Engg. & Tech., Chennai, Tamilnadu, India [1]

Assistant professor, Dept. of ECE, Sree Sastha Institute of Engg. & Tech., Chennai, Tamilnadu, India [2]

ABSTRACT: Speech and music are the most basic means of human communication. As technology advances and increasingly sophisticated tools become available to extract human speech from a noisy background. But the task of extracting a singing voice from a musical background, composed of many musical instruments is a challenging one as both the signals have very high coherence and correlation. Separating singing voice from music accompaniment is very useful in many applications, such as lyrics recognition and alignment, singer identification, and music information retrieval. This paper describes, a trend estimation algorithm to detect the pitch ranges of a singing voice in each time frame. The detected trend substantially reduces the difficulty of singing pitch detection by removing a large number of wrong pitch candidates either produced by musical instruments or the overtones of the singing voice. Qualitative results show that the system performs the separation task successfully.

KEYWORDS: Trend estimation, pitch detection, singing voice separation.

I.INTRODUCTION

Singing voice separation is, in a sense, a special case of speech separation and has many similar applications. For example, automatic speech recognition corresponds to automatic lyrics recognition, automatic speaker identification to automatic singer identification, and automatic subtitle alignment which aligns speech and subtitle to automatic lyric alignment which can be used in a karaoke system. Compared to speech separation, separation of singing voice could be simpler with less pitch variation. On the other hand, there are several major differences. For speech separation, or the cocktail party problem, the goal is to separate the target speech from various types of background noise which can be broadband or narrowband, periodic or a periodic. In addition, the background noise is independent of speech in most cases so that their spectral contents are uncorrelated. For singing voice separation, the goal is to separate singing voice from music accompaniments which in most cases are broadband, periodic, and strongly correlated to the singing voice. Furthermore, the upper pitch boundary of singing can be as high as 1400 Hz for soprano singers while the pitch range of normal speech is between 80 and 500 Hz. These differences make the separation of singing voice and music accompaniment potentially more challenging.

The singing voice separation, existing methods can be generally classified into three categories depending on their underlying methodologies: spectrogram factorization, model-based methods and pitch–based methods. Spectrogram factorization methods utilize the redundancy of the singing voice and music accompaniment by decomposing the input signal into a pool of repetitive components. Each component is then assigned to a sound source. Model-based methods learn a set of spectra from music accompaniment only segments. Spectra of the vocal signal are then learned from the sound mixture by fixing accompaniment spectra. Pitch-based methods use extracted vocal pitch contours as the cue to separate the harmonic structure of the singing voice. Musical sound separation systems attempt to separate individual musical sources from sound mixtures. The human auditory system gives us the extraordinary capability of identifying instruments being played (pitched and Non-pitched) from a piece of music and also hearing the rhythm/melody of the individual instrument being played. This task appears 'automatic' to us but has proved to be very difficult to replicate in computational systems.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 2, April 2014

Hu and Wang [1] proposed a tandem algorithm which performs pitch estimation and voice separation jointly and iteratively. The tandem algorithm gives more than one pitch candidate for each frame and has the problem of sequential grouping (ie., deciding which pitch contour belongs to the target). Wang and Brown **[**9**],** proposed a new channel/peak selection scheme to exploit the salience of singing voice and the beating phenomenon in high frequency channels. An HMM model is employed to integrate the periodicity information across frequency channels and time frames which improves the accuracy of predominant pitch detection for singing voice. The problem is that the low frequency channels do not provide enough information in distinguishing different sound sources in the presence of strong percussive sounds as encountered in country and rock music.

Klapuri et al [7], focused on the problem of identifying segments of singing within popular music as a useful and tractable form of content analysis for music, particularly as a precursor to automatic transcription of lyrics. In [2], pitch is detected based on a hidden Markov model (HMM). Here, a predominant pitch detection algorithm is proposed which can detect the pitch of singing voice for different musical genres even when the accompaniment is strong. One problem with this approach is that the frequency resolution in the high-frequency range is limited. As a result this system cannot be used to separate high-pitched singing voice. However, most types of singing, such as in pop, rock, and country music, have a smaller pitch range and, therefore, this system can potentially be applied to a wide range of problems.

Wang and Brown [5] proposed a robust algorithm for multipitch tracking of noisy speech. This approach incorporates Pitch Determination Algorithms (PDAs) for extracting periodicity information across different channels and a Hidden Markov Model (HMM) for continuous pitch tracks. A common problem in PDAs is harmonic and subharmonic errors, in which the harmonics or subharmonics of a pitch are detected instead of the real pitch itself. Here, the performance drops significantly when the number musical instruments increases.

II.SYSTEM DESCRIPTION

Our system consists of three stages. The input to the system is a mixture of singing voice and music accompaniment. In the singing voice detection stage, the input is first partitioned into spectrally homogeneous portions by detecting significant spectral changes. Then, each portion is classified as a vocal portion in which singing voice is present, or a nonvocal portion in which singing voice is absent.

The predominant pitch detection stage detects the pitch contours of singing voice for vocal portions. In this stage, a vocal portion is first processed by a filterbank which simulates the frequency decomposition of the auditory periphery. After auditory filtering periodicity information is extracted from the output of each frequency channel. Next a hidden Markov model (HMM) is used to model the pitch generation process. Finally, the most probable pitch hypothesis sequence is identified as pitch contours of the singing voice using the Viterbi algorithm.

The separation stage has two main steps: the segmentation step and the grouping step. In the segmentation step, a vocal portion is decomposed into T-F units, from which segments are formed based on temporal continuity and cross-channel correlation. In the grouping step, T-F units are labeled as singing dominant or accompaniment dominant using detected pitch contours. Segments in which the majority of T-F units are labeled as singing dominant are grouped to form the foreground stream, which corresponds to singing voice. Separated singing voice is then resynthesized from the segments belonging to the foreground stream. The output of the overall system is the separated singing voice.

The following subsections explain each stage in detail.

A. Singing Voice Detection

The goal of this stage is to partition the input into vocal and nonvocal portions. Therefore, this stage needs to address the classification and partition problem. For the classification problem, the two key components in the system design are features and classifiers.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 2, April 2014

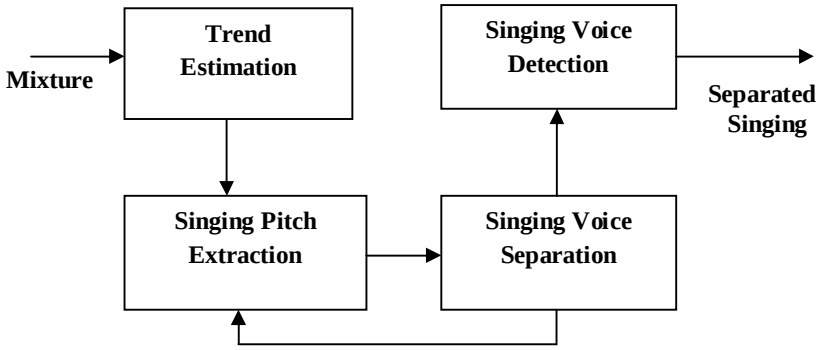


Fig. 1 Schematic diagram of the proposed system.

when a new sound enters a mixture, it usually introduces significant spectral changes. As a result, the possible instances of a sound event in a mixture can be determined by identifying significant spectral changes.

The spectral change detector calculates the Euclidian distance η(m) in the complex domain between the expected spectral value and the observed one in a frame (A frame is a block of samples within which the signal is assumed to be near stationary).

$$\eta(m) = \sum_k \left(\left| \hat{S_k}(m) - S_k(m) \right| \right) \quad (1)$$

where $S_k(m)$ is the observed spectral value at frame m and frequency bin k. $\hat{S_k}$(m) is the expected spectral value of the same frame and the same bin, calculated by

$$\hat{S_K}(m) = |S_k(m-1)| e^{j\hat{\emptyset_k}(m)} \quad (2)$$

where $|S_k(m-1)|$ is the spectral magnitude of the previous frame at bin k. $\hat{\emptyset_k}$(m) is the expected phase which can be calculated as the sum of the phase of previous frame and the phase difference between the previous two frames.

$$\hat{\emptyset_k}(m) = \emptyset_k(m-1) + \left(\emptyset_k(m-1) - \emptyset_k(m-2) \right) \quad (3)$$

where $\emptyset_k$(m-1) and $\emptyset_k$(m-2) are the unwrapped phases for frame m-1 and frame m-2 respectively. η(m) is calculated for each frame of 16 ms with a frame shift of 10 ms.

A local peak in η(m) indicates a spectral change, which can either be that the spectral contents of a sound are changing or a new sound is entering the scene. To accommodate the dynamic range of the spectral change as well as spectral fluctuations, dynamic thresholding is applied to identify the instances of significant spectral changes. Specifically, a frame m will be recognized as an instance of significant spectral change if η(m) is greater than the weighted median value in a window of size H=10

$$\eta(m) > C \times median\left(\eta\left(m - \frac{H}{2}\right), \ldots, \eta\left(m + \frac{H}{2}\right) \right) \quad (4)$$

where C=1.5 corresponds to the weighting factor.

By made use of 13 triangular filters in the filter bank and thus generated 13 MFCC coefficients per frame. Finally, the mel-frequency cepstral coefficients (MFCCs) coefficients are used as the short- term feature for classification and are

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 2, April 2014

calculated for all the frames. The portion between two consecutive spectral change instances is relatively homogeneous, and the short-term classification results can then be pooled over the portion to yield more reliable classification. Now Gaussian mixture models (GMMs) is used to classify a frame as belonging to one of the two clusters (vocal or non-vocal).

B. Predominant Pitch Detection.

In the second stage, portions classified as vocal are used as input to a predominant pitch detection algorithm. Our predominant pitch detection starts with an auditory peripheral model for frequency decomposition. The signal is sampled at 16 kHz and passed through a 64-channel gammatone filterbank. The centre frequencies of the channels are equally distributed on the equivalent rectangular bandwidth (ERB) scale between 80 Hz and 5 kHz.

The first stage output once decomposed into channels using the filter bank, is split up into frames for a duration of 20 ms with 10 ms overlap on either side. Thus a single frame belonging to a channel is said to be a Time-Frequency unit or T-F unit. Let u_{cm} denote a T- F unit at channel c and frame m, and y(c, t) the filtered signal at channel c and time t. The corresponding normalized correlogram A(c, m, τ) at u_{cm} is computed by the following autocorrelation function (ACF):

$$A(c,m,\tau) = \frac{\sum_n y(c, mT_m - nT_n)\, y(c, mT_m - nT_n - \tau T_n)}{\sqrt{\sum_n y^2\,(c, mT_m - nT_n) \sum_n y^2 (c, mT_m - nT_n - \tau T_n)}} \quad (5)$$

where τ is the time delay. T_m is the frame shift and T_n is the sampling time. The above summation is over 40 ms, the length of a time frame. The peaks of the ACF indicate the periodicity of the filter response, and the corresponding delays indicate the periods.

The cross- channel correlation measures the similarity between the responses of two adjacent filters, indicate whether the filters are responding to the same sound component. Hence, we calculate the cross - channel correlation between u_{cm} and $u_{c+1,m}$ by,

$$C(c,m) = \frac{\sum_\tau [A(c,m,\tau) - \bar{A}(c,m)][A(c+1,m,\tau) - \bar{A}(c+1,m)]}{\sqrt{\sum_\tau [A(c,m,\tau) - \bar{A}(c,m)]^2 \sum_\tau [A(c+1,m,\tau) - \bar{A}(c+1,m)]^2}} \quad (6)$$

where $\bar{A}$ denotes the average of A over τ.

Channels with centre frequencies above 800Hz are treated as high frequency channels. For the listed high frequency channels, Teager energy operator and a low- pass filter are used to extract the envelopes in high frequency channels. For a digital signal S_n, the Teager energy operator is defined as,

$$E_n = S_n^2 - S_{n+1} S_{n-1} \quad (7)$$

Then the signals are low- pass filtered at 800 Hz using the third- order Butterworth filter. The corresponding output is subjected to correlogram and replaces the original correlogram for high frequency channels.

To detect the dominant pitch belonging to a frame, all the channel outputs are summed up and normalized. The first peak occurring within a duration of 2.5- 12.5 ms (80- 400 Hz is the human being's pitch range) with a threshold greater

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 2, April 2014

than 0.6 is taken to be the predominant pitch period. If no such result is obtained in a particular frame, it is said to be music-dominant.

Our tandem algorithm detects multiple pitch contours and separates the singer by estimating the ideal binary mask (IBM) which is a binary matrix constructed using premixed source signals. In the IBM, 1 indicates that the singing voice is stronger than interference in the corresponding time-frequency unit and 0 otherwise. A T- F unit is labeled 1 if and only if the corresponding response or response envelope has a periodicity similar to that of the target.

C. Singing Voice Separation

The separation stage has two main steps: the segmentation step and the grouping step. In the segmentation step, In the segmentation step, our algorithm extracts following features for each T-F unit: energy, autocorrelation, cross-channel correlation, and cross-frame correlation. Next, segments are formed by merging contiguous T-F units based on temporal continuity and cross-channel correlation. Only those T-F units with significant energy and high cross-channel correlation are considered.

In the grouping step, the Trend estimation algorithm applies an iterative method to estimate the pitch contours of the target signal. Since we have already obtained predominant pitch contours, we directly supply detected pitch contours in the grouping step. A T-F unit is labeled as singing dominant if its local periodicity matches the detected pitch point of the frame. If the majority of the T-F units within a certain frame are labeled as singing dominant, the segment is said to be dominated by singing voice at this frame. All the singing dominant segments are grouped to form the foreground stream, which corresponds to the singing voice.

III. EVALUATION

In this section, we evaluate the performance of the whole separation system.

A famous song for a duration 20 seconds has been fed as the input to the system as shown in Fig. 2. Identification of spectral changes between frames is necessary because when a new sound enters or leaves a mixture, it usually introduces significant spectral changes. As a result, the possible instances of a sound event in a mixture can be determined by identifying significant spectral changes. The Fig. 3 shows the spectral changes between adjacent frames for the input song.

At the end of the singing voice detection stage, frames where music is alone present are removed by applying the mask which is obtained by combining instant spectral changes and Gaussian Mixture Model (GMM) outputs. But the portions where both vocal and music occur at the same time have not yet been removed. The Fig. 4. shows the output of the first stage.

The output of the first stage is fed to a gammatone filter bank with 64 filters. The output is then divided into frames with size 20ms and 10 ms overlap on either side and individual frame correlation for every channel also known as correlogram is computed. Using this, the predominant pitch for each frame is estimated. The mask obtained due to cross- channel and cross-frame correlation are shown in Fig. 5(a) and Fig. 5(b) respectively. The mask combining cross-channel and cross-frame correlation is shown in Fig. 5(c). The final binary mask where 1 indicates that the singing voice is stronger than the interference in the corresponding time-frequency unit and 0 otherwise is shown in Fig. 6. The final vocal-only output portions obtained for the input song is illustarted in Fig. 7.

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 2, April 2014

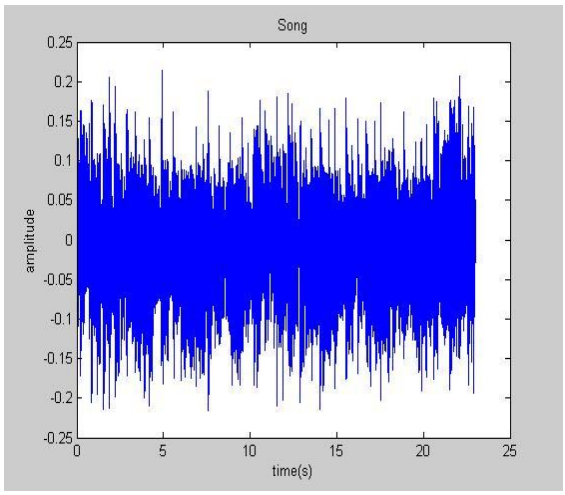


Fig. 2 The input song mixture of duration 20 seconds.

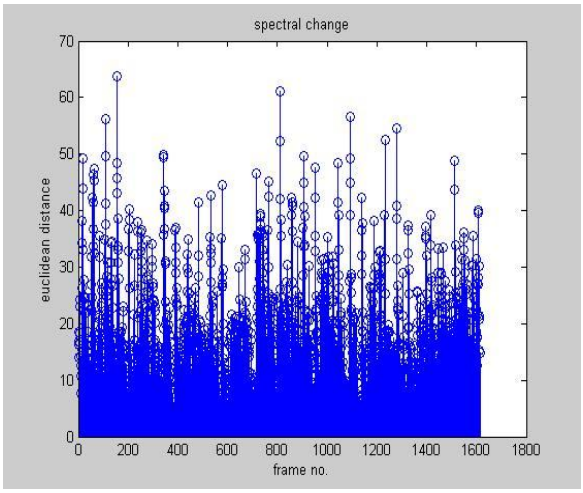


Fig. 3 Magnitude of spectral differences between successive frames.

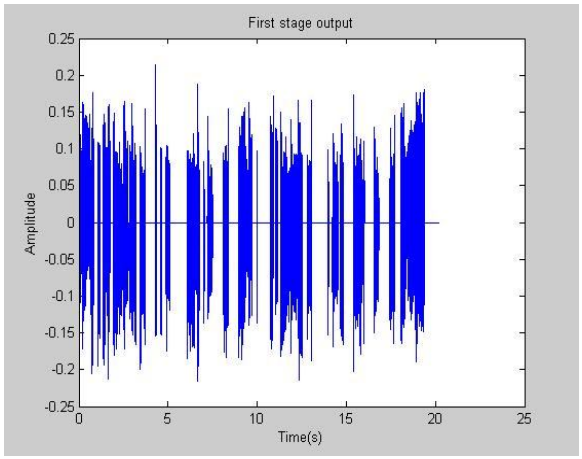


Fig. 4 The output of the singing voice detection stage.

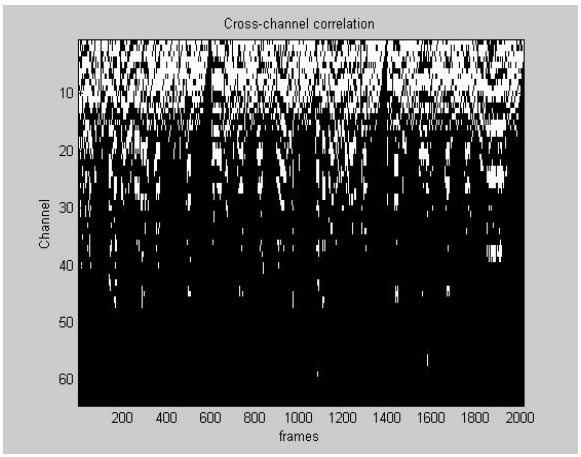


Fig. 5(a) Resultant mask due to cross-channel correlation

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 2, April 2014

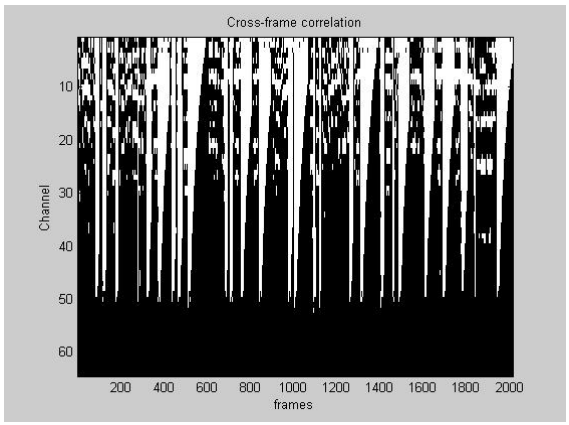


Fig. 5(b) Resultant mask due to cross-frame correlation.

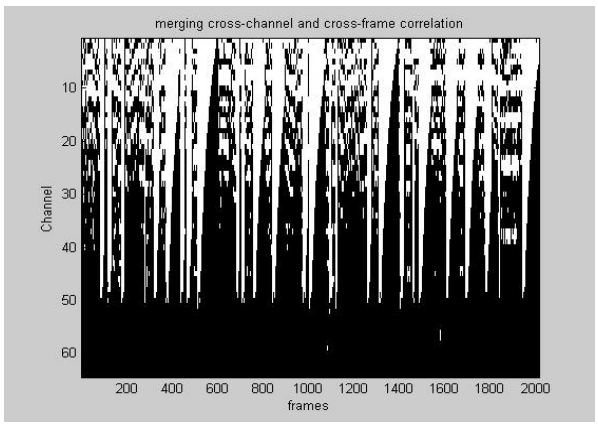
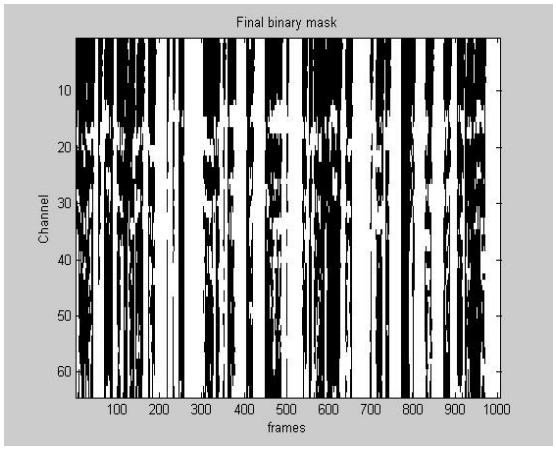


Fig. 5(c) Mask combining cross-channel and cross-frame correlation.



.
Fig. 6. Final T-F binary mask.

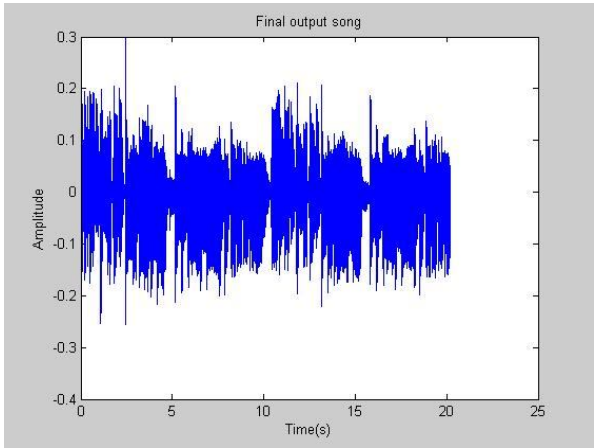


Fig. 7. Final output containing voice only.

IV.CONCLUSION

As mentioned in the Introduction, few systems have been proposed for singing voice separation. By integrating singing voice classification, predominant pitch detection, and pitch-based separation, our system represents the first general framework for singing voice separation. Another important aspect of the proposed system is its adaptability to different genres. Currently, our system is genre independent, i.e., rock music, carnatic music, cine music and country music are treated in the same way. This, in a sense, is strength of the proposed system. However, considering the vast variety of music, a genre-dependent system may achieve better performance. Given the genre information, the system can be adapted to the specific genre. The singing voice detection stage can be retrained using genre-specific samples. We can also extend our algorithm to applications such as Singing voice recognition, Lyrics recognition, Language identification, Song remix, Male and female voice separation, Karaoke application, Converting male voice into female voice and vice-versa

ISSN (Print) : 2320 – 3765
ISSN (Online): 2278 – 8875

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 2, April 2014

We have demonstrated an example for pitch translation from female to male voice by using voice-only portions of the song. Final remixing was done with the original music. We are looking forward to apply this algorithm for all the listed applications.

REFERENCES

[1] G. Hu and D.L. Wang, "A tandem algorithm for pitch estimation and voiced speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067- 2079, Nov. 2010.
[2] Y. Li and D.L. Wang, "Separation of singing voice from music accompaniment for monaural recordings," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no.4, pp. 1475-1487, May 2007.
[3] Zhaozhang Jin and DeLiang Wang, "HMM- based multipitch tracking for noisy and reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 5, pp. 1091- 1102, July 2011.
[4] G. Hu and D.L Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135- 1150, Sep. 2004.
[5] M. Wu, D.L. Wang and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no.3, pp. 229-241, May 2003.
[6] G.J. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, pp. 297- 336, 1994.
[7] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 204- 816, Nov 2003.
[8] Y. Li and D.L. Wang, "Detecting pitch of singing voice in polyphonic audio," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2005, vol. 3, pp. 17-20.
[9] D. Wang and G. Brown, "An auditory scene analysis approach to monaural speech segregation," in Topics in Acoustic Echo and Noise Control, E. Hansler and G. Schmidt, Eds. Heidelberg, Germany: Springer, pp. 485-515, 2006.
[10] Video lectures on Speech Processing by Prof. E. Ambikairajah, University of New South Wales, *http://www.onlinevideolectures.com/the-university-of-new-south-walves*.
[11] Lawrence Rabiner and Biing- Hwang Juang, "Fundamentals of speech recognition," 4th Edition, Prentice Hall, 1978.