IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Tandem algorithm with Supervised Classifier for Pitch Estimation and Voice Separation from Music Accompaniments

Nichal Vikas R[1], Mane Vikram A[2]

Department of E&TC, Annasaheb Dange COE, Asha, Maharashtra, India[1]

Department of E&TC, Annasaheb Dange COE, Asha, Maharashtra, India[2]

Abstract: Singing voice separation from music is a kind of speech separation and it is a big challenge in many applications. In this paper, support vector machine with tandem algorithm is proposed to estimate the singing pitch and separate the singing voice and music from music accompaniments. Detecting the pitch range of the singing voice and minimizing the spurious pitches occurring due to higher order harmonics are done by trend estimation algorithm. In tandem algorithm, the pitch is estimated first and then the multiple pitch contours and their associated time-frequency masks are obtained. Then the pitch contours are expanded according to temporal continuity. A post-processing stage is introduced to deal with the "sequential grouping" problem. Once tandem algorithm detects multiple pitch contours, the nest stage separates the singing voice by estimating the ideal binary mask (IBM), which is a binary matrix, constructed using premixed source signals. This stage employs a continuous SVM to decode an input mixture into vocal and nonvocal sections. Finally, separated voice is used to extract music from the mixture signal. The experimentation is performed using a signal containing voice and music, and the performance is evaluated using precision, recall and accuracy.

Keywords: Support vector machine, music accompaniments, Pitch, Trend estimation, Tandem algorithm.

I. INTRODUCTION

The music databases both with professional and personal requirements have rapidly grown because of popularization and wide usage of digital music. The trending of technologies that deal with categorization and retrieval has also risen in response to the requirements and consumer demands. The automatic singer voice extraction technology not only acts as an application, but also working with various applications and acts as sub-processes [1].

The necessity of such technology has been extended to a wide end. This technology intends to extract a particular singer's voice from music accompaniments based on certain feature sets like pitch[2,3].Singing voice separation is a special kind of speech separation [14,17].

The aim of speech separation is to separate the target speech from broadband or narrowband, periodic or aperiodic background. Where As, for singing voice separation, separating singing voice from broadband, periodic and correlatedmusic accompaniments are the major goals. Also, another challenge is the difference in the pitch range of singers (approximately 1400 Hz) and the normal speech (between 80 and 500 Hz) [1,5].

There are different traditional methods developed for singing voice separation from music like the singing voice separation by using a harmonic-locked loop technique. As this system needs the estimation of a partials instantaneous frequency, the system cans only works in conditions where the singing voice to accompaniments energy ratio is high [18].

The Monaural speech segregation is a technique is based on pitch tracking and amplitude modulation.

Here, the estimation of pitch is unreliable for singing voice. So, for singing voice separation, this system cannot separate unvoiced speech [2]. In Computational auditory scene analysis (CASA) method a lot of effort has been made to segregate speech from music accompaniments. But pitch estimation errors and residual noise reduces the performance of the system [8,9, 10, 13].

In this paper, we propose support vector machine with tandem algorithm and to estimate the singing pitch and separate the singing voice and music from music accompaniments. Detecting the pitch range of the singing voice and reducing the false pitches due to higher order harmonics are done by trend estimation algorithm. In tandem algorithm, the pitch is estimated and the multiple pitch contours and their associated time-frequency masks is obtained. A post-processing stage is introduced to deal with the "sequential grouping" problem. Once tandem algorithm detects multiple pitch contours, the nest stage separates the singing voice by estimating the ideal binary mask (IBM), which is a binary matrix, constructed using premixed source signals. This stage employs a continuous SVM to decode an input mixture into vocal and nonvocal sections. Finally, separated voice is used to extract music from the mixture signal. The proposed support vector machine with tandem Algorithm for voice and music separation is given in section 2. In section 3, comparison

DOI 10.17148/IJARCCE.2016.5617

of SVM method with HMM (Hidden Markov Model) [4] is presented. Conclusion is presented in section 4.

II. PROPOSED SUPPORT VECTOR MACHINE WITH TANDEM ALGORITHM FOR VOICE AND MUSIC SEPARATION

The proposed system separates the voice signal from the music signal as per the proposed diagram shown in figure 1. It has three major parts namely, Trend estimation, Tandem algorithm, andVoice extraction by SVM. Spurious pitches arising due to music accompaniments or due to higher order harmonics are reduced by trend estimation. In tandem algorithm [7], an iterative processing is performed which considers mask estimation and pitch detection in a sequence.

The initial estimation is done through harmonic/percussive source separation (HPSS) process. Ideal Binary Mask (IBM) is estimated in the mask estimation process. The pitch is then determined based on the estimated binary mask. For detecting the pitch process, recursive operation for a number of iterations is performed in tandem algorithm. [16].

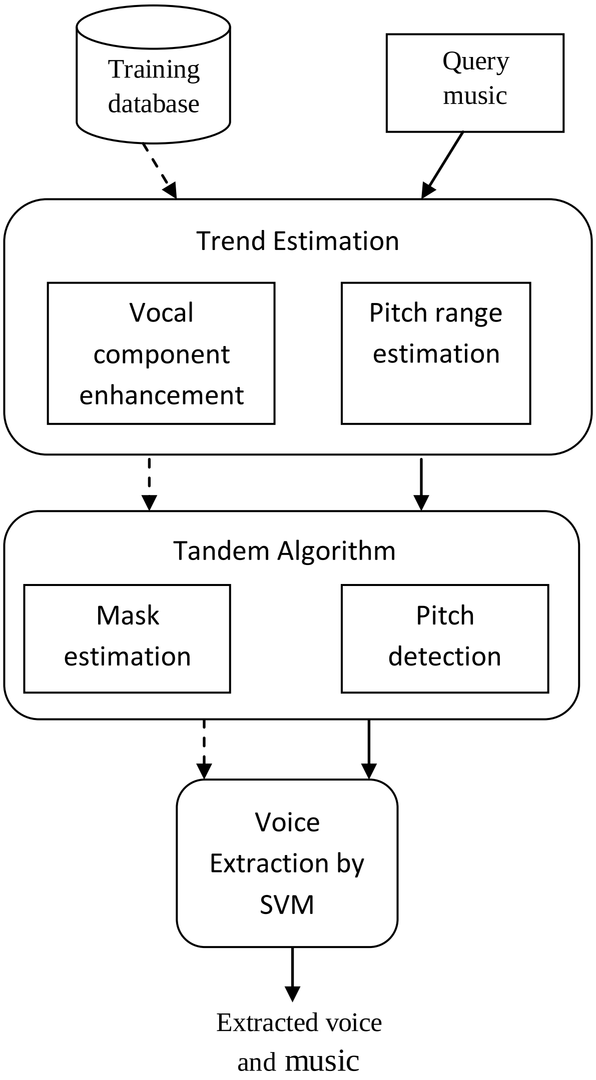


Figure 1. Overview of the proposed methodology for pitch estimation and singer voice separation

After pitch extraction, the detected pitch is subjected to supervised voice extraction process. This work uses Support Vector Machine (SVM) to perform voice extraction process. Its supervisory nature enhances the performance of voice detection performance.

A. Design of Trend estimation

Trend estimation algorithm is explained in this section.The purpose of trend estimation is to find a sequence of relatively tight pitch ranges where the F0s of the singing voice are present. This is achieved by removing unreliable peaks and higher harmonics of the singing voice. The peaks are bounded in a sequence of T-F blocks, and the F0 range is estimated [6,15].

Multi-resolution fast Fourier transform (MR-FFT) proposed by Dressler [12] is applied which deletes the unreliable peaks based on the local sinusoidality criteria. The instantaneous frequencies of the peaks in each time frame are checked. If their instantaneous frequencies are

DOI 10.17148/IJARCCE.2016.5617

less than 0.2 semitones, the frequency with the largest magnitude is selected. Then the harmonics are deleted to make the vocal F0 be the only lowest frequency within a frame[16,17].

The pitch range of singing voice is then estimated by using the remaining peaks.For a multi-resolution spectrogram $x[d, f]$ produced by MR-FFT, the downsampled magnitude of the Time-Frequency block structure is given as

$$B(T,F) = \sum_{d=0}^{D_T-1} \max_{f\in[0,D_F-1]} x[d+TL_{T,}\ f+FL_f],$$
$$T = 0,1,.......R_1 - 1 \ and \ F = 0,1,,,,,,,R_2 - 1 \quad (1)$$

Where T and F indicate the time and frequency indices of a Time-Frequency block, respectively. R_1 and R_2 indicate the ranges of Time-Frequency blocks in time and frequency, respectively. D_T and D_F are the numbers of time frames and frequency bins of a T-F block, respectively, and L_T and L_F are the shifts in frame and frequency bins of a Time-Frequency block, respectively.

Finally, using dynamic programming, an optimal path consisting of a sequence of Time-Frequency blocks is found that contain the largest magnitudes. The problem is defined as finding an optimal path $[F_0, F_1,........F_1]$ that maximizes the following score:

$$\sum_{T=0}^{P=1} B(T, F_T) - \theta \sum_{T=0}^{P-1} |F_T - F_{T-1}| \quad (2)$$

where the first term is the sum of strengths of the T-F blocks along the path, and the second term controls the smoothness of the path. For larger value of θ , the computed path will be very smooth.The Viterbi algorithm [4] is used to find the optimal path.

The upper boundary and the lower boundary of the selected T-F block is extended and the estimated trend is formed. This makes trend estimation to tolerate the possible pitch changes of the singing voice.

B. Tandem algorithm for pitch estimation

a) Initial estimation of pitch

In pitch estimation, the output from HPSS is used as the input signal [11]. First all T-F units with high cross-channel correlations are treated as dominated by a single source. Then the estimated pitch period is selected as the one supported by most active (value 1) T-F units. A T-F unit u_{cd} is considered supporting a pitch period τ if the corresponding $P(H_0|r_{cd}(\tau))$ is higher than 0.75. The T-F units that do not meet the threshold are then used to estimate the second pitch period if such units exist. The possible pitch periods are now confined to the pitch range of the estimated trend. With the estimated pitch period, the corresponding mask is reestimated as the target if

$$P(H_0|r_{cd}(\tau)) > 0.5 .$$

After the above estimation, individual pitch periods are combined into pitch contours based on temporal continuity of both pitch periods and corresponding masks. As a result of this step, multiple pitch contours and their associated T-F masks is obtained.

b) Iterative estimation of pitch

In this step, the pitch contours are expanded according to temporal continuity. Let p_k be a pitch contour containing a sequence of pitch points in a continuous set of frames and $L_k(d) = \{L_k(c,d), \forall c\}$ be the associated mask at frame m . First expand the mask by letting $L_k(d_1 - 1) = L_k(d_1)$ and $L_k(d_2 + 1) = L_k(d_2)$, where d_1 and d_2 are the first and last frame of the pitch contour, respectively.

A new p_k is then estimated from the new mask [2]. If the newly estimated pitch points pass the continuity criterion [2], it is considered as reliable pitch. Otherwise, it is discarded.The above two steps iterate until the estimation of pitch and mask converges.

c) Remove pitch out of the plausible range

For each frame, two pitch values are obtained as output from the pitch estimation algorithm.So, some pitch contours may overlap each other in time. This creates a sequential grouping problem.

The proposed trend estimation algorithm is able to remove most of the time-overlapped pitch contours. This makes the sequential grouping problem easy to solve.

d) Produce a pitch contour matrix

Pitch contours are grouped as follows. First, pitch points are assigned to the target pitch track for the frames that have only one pitch candidate.

For the frames that contain more than one pitch candidate, the pitch that is supported by most channels is selected as the target pitch. Then, the gaps where no reliable pitch is estimated is filled by using the most dominant pitch period in the initial estimation and thus generate a continuous pitch track.

e) Create overall pitch and mask

Two key steps of our tandem algorithm are IBM estimation given target pitch and pitch determination given a binary mask.

i) IBM estimation given target pitch

After generating the pitch counter matrix, the corresponding mask to the target pitch track forms the estimated IBM.

DOI 10.17148/IJARCCE.2016.5617

ii) Pitch Estimation Given Binary Mask
The target pitch is estimated by summing the autocorrelations across all the channels and then identifying the most dominant peak in the summary correlogram. This can be improved by calculating the summary correlogram only from target-dominant T-F units according to the given binary mask $L(c,d)$.

The value of 1 indicates that u_{cd} is dominated by the target and 0 otherwise. Also, as shown in [2], replacing ACF with $P(H_0|r_{cd}(\tau))$ improves pitch estimation and the pitch period is estimated by

$$SP_d(\tau)=\sum_c P(H_0|r_{cd}(\tau))L(c,d) \qquad (3)$$

Since results of one stage are inputs to the other one, the two stages are used to improve each other iteratively.

C. Singing Voice And Music Detection Through The Estimated Pitch And Mask By Svm Continuous SVM is used to separate an input mixture signal into vocal and nonvocal signals. The signals after applying HPSS attenuate the energy from music accompaniment instead of the original mixture. SVM generally outperforms methods, such as HMM for speech detection.
The SVM maximize the distance between two samples. SVM is used for a linearly nonseparable scenario to perform separation. The SVM is trained by minimizing the following cost function

$$f(\omega,\xi)=\|\omega\|^2/2+C\sum_i \xi_i \qquad (4)$$

with the constraints

$$y_i\left(w'\varphi(z_i)+b\right)\geq 1-\xi_i\,; \qquad \xi_i\geq 0$$

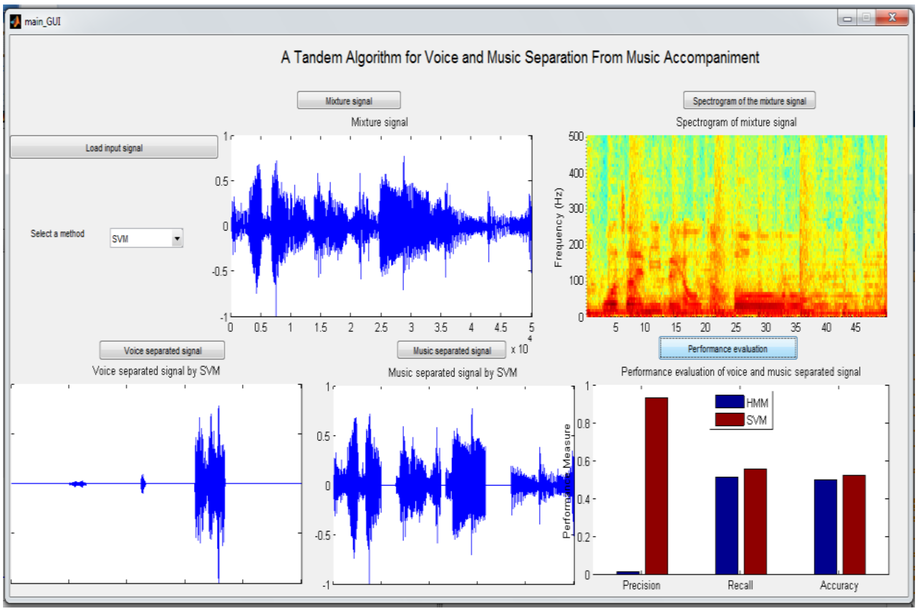
w represents the weight vector of the separating hyperplane and ξi is a nonnegative slack variable corresponding to the classification error. C controls the tradeoff between the margin of two classes and the separation errors.

Φ is a mapping function which projects the training features to a higher-dimensional space, yi is the label for zi, and b is the bias. Transpose is denoted by‘. However, in our case, the two classes of training samples are unbalanced. This may cause the SVM hyperplane to skew to the minority class.
To compensate this imbalance, development set is used to search for a threshold to binarize SVM outputs to maximize the classification accuracy for each channel. The new thresholds are used for binarization of SVM outputs in testing.
By combining all unvoiced-dominant T-F units in UV intervals, mask corresponding to the segregated unvoiced speech is obtained.

III. RESULTS AND DISCUSSION

In this section the performance of tandem algorithm with SVM and tandem algorithm with HMM for singing pitch extraction and voice separation from music accompaniment is evaluated.
Here, we use a signal which contains both voice and music to evaluate our proposed system. The performance of singing voice detection is represented by precision, recall and overall accuracy.

A.GUI of the proposed method
The Graphical User Interface of the proposed method is given in the figure 2. The input signal is a mixture signal. The spectrogram of the mixture signal is also shown in the GUI. By selecting the SVM method, the voice separated by SVM and the Music separated by SVM is obtained. Finally, the performance of the voice and music signal separation by HMM and SVM is compared.


main_GUI

A Tandem Algorithm for Voice and Music Separation From Music Accompaniment

Select a method: SVM

Figure.2 GUI of the proposed method

B. Experimental results

The input of the proposed system is mixture signal. It is shown in figure 3. The spectrogram of the mixture signal and the voice and music separated signals are given in the following figures. The output of the signals taken for the experimentation is given in figure 5 to 8. Figure.9show the performance graph of the proposed SVM model and HMM model.

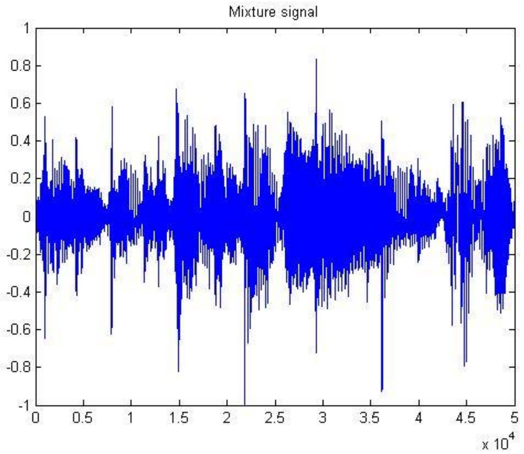


Figure. 3 Mixture signal

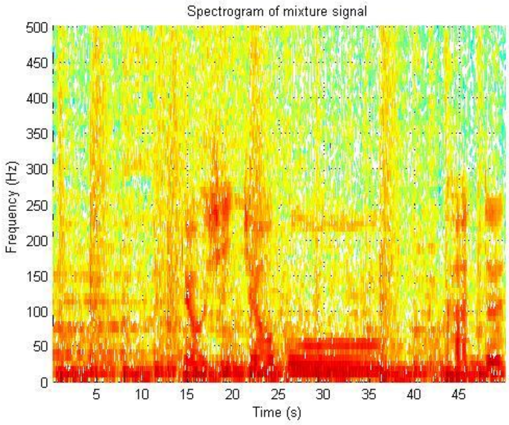


Figure. 4 Spectrogram of mixture signal

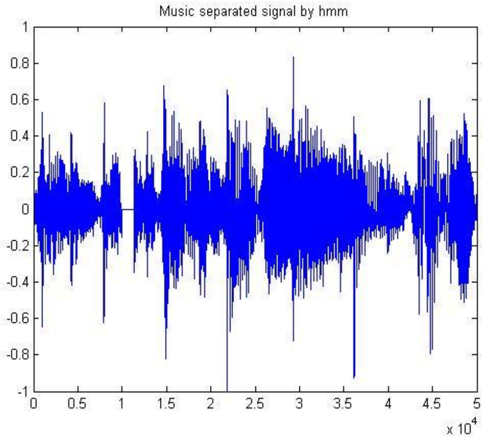


Figure. 5 Music separated by HMM

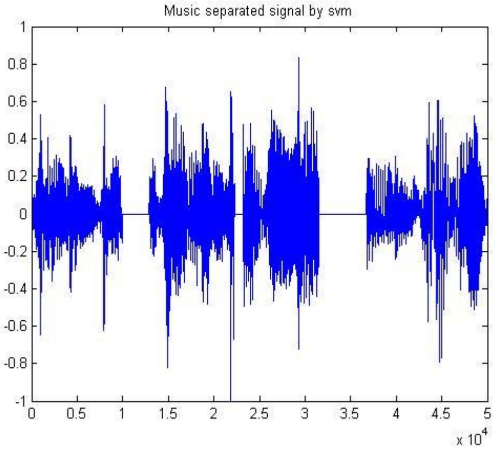


Figure. 6 Music separated by SVM

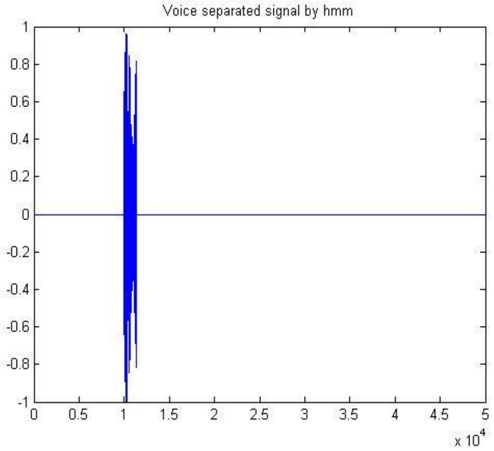


Figure. 7 Voice separated by HMM

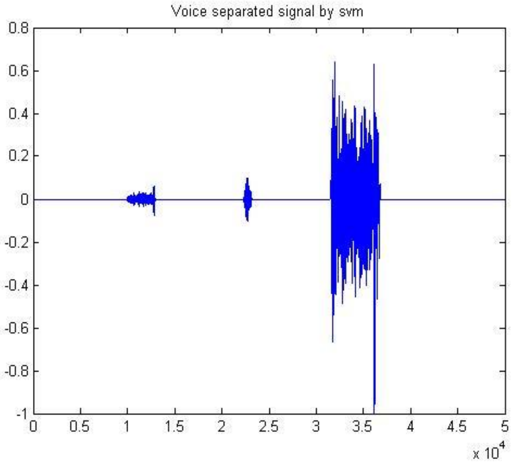


Figure. 8 Voice separated by SVM

C. Performance evaluation

From the performance graph, it is noted that the proposed SVM method outperforms HMM for voice and music separation.

DOI 10.17148/IJARCCE.2016.5617

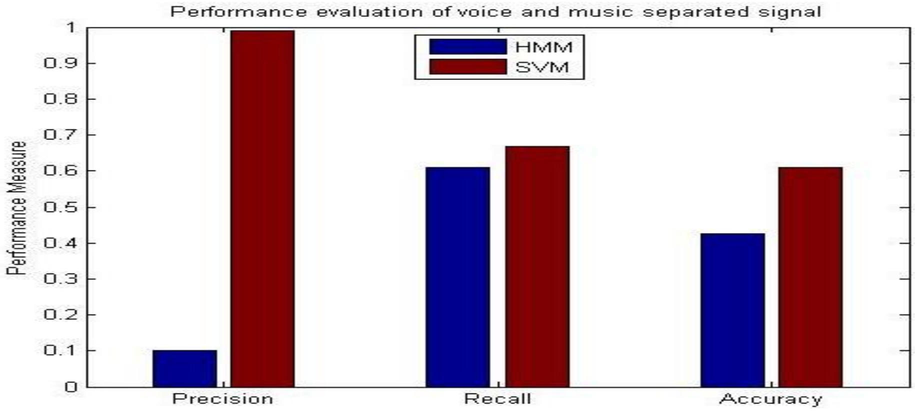


Figure.9 Performance evaluation of voice and music separated signal

4. CONCLUSION

In this work, tandem and SVM is used for voice and music separation. Here, the tandem algorithm is investigated and extended to separate the voiced portions of singing voice from music accompaniment. Trend estimation algorithm first estimates the pitch ranges of the singing voice. The estimated trend is then incorporated in the tandem algorithm to acquire the initial estimate of the singing pitch. Singing voice is then separated according to the initially estimated pitch. 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. Finally, singing voice detection is performed to discard the nonvocal parts of the separated singing voice. Future work needs to analyze the tradeoff to see if there is an optimal trend range for pitch detection. Also, intelligence classifier can be utilized instead of SVM for further improving the estimation.

REFERENCES

[1] Chao-Ling Hsu, DeLiang Wang, Jyh-Shing Roger Jang, and Ke Hu, “A Tandem Algorithm For Singing Pitch Extraction and Voice Separation from Music Accompaniment”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 5, p.p. 1482-1491, 2012.

[2] 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.

[3] Y. Meron and K. Hirose, “Separation of singing and piano sounds,” in Proc. 5th Int. Conf. Spoken Lang. Process. (ICSLP 98), 1998.

[4] L. R. Rabiner, “A tutorial on hidden Markov models and selected application in speech recognition,” Proc. IEEE, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[5] T. Zhang, “Perception of singing,” in Psychology of Music, D. Deutsch, Ed., 2nd ed. New York: Academic, 1999, pp. 171–214.

[6] M. Goto, “A real-time music-scene-description system: Predomin ant- F0 estimation for detecting melody and bass lines in real-world audio signals,” Speech Commun., vol. 43, no. 4, pp. 311–329, 2004.

[7] Guoning Hu and Deleing Wang “A Tandem algorithm for pitch estimation and voice speech Segregation” IEEE Transaction on Audio, Speech, and Language Processing, VOL.18, NO.8, NOVEMBER 2010.

[8] A. Bregman, Auditory Scene Analysis. Cambridge, MA: MIT Press, 1990.

[9] D. Wang and G. Brown, Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Hoboken, NJ: Wiley and IEEE Press, 2006.

[10] D. L. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in Speech Separation by Humans and Machines,P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.

[11] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, “Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram,” in Proc. EUSIPCO, 2008.

[12] K. Dressler, “Sinusoidal extraction uses an efficient implementation of a multi-resolution FFT,” in Proc. Int. Conf. Digit. Audio Effects, 2006. pp. 247–252.

[13] G. J. Brown and M. Cooke, “Computational auditory scene analysis,” Comput. Speech Lang., vol. 8, pp. 297–336, 1994.

[14] D. L. Wang and G. J. Brown, “Separation of speech from interfering sounds based on oscillatory correlation,” IEEE Trans. Neural Netw., vol. 10, no. 3, pp. 684–697, May 1999.

[15] A. de Cheveigne, “Multiple F0 Estimation,” in Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley and IEEE Press, 2006, pp. 45–79.

[16] G. Hu and D. L. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[17] 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.

[18] A.L.C Wang. “Instantaneous and frequency-warped signal processing Technique for auditory Source separation.” Ph.D. dissertation, Dept. Elect. Eng., Stanford Univ., Stanford, CA, 1994.

[19] 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.

BIOGRAPHIES

Vikas Nichal received the BE degree in Electronics and Tele-communication engineering from Shivaji university, Kolhapur in 2012. He now is doing ME in Electronics and Tele-communication engineering from Shivaji university, Kolhapur. His area of specialization in digital signal processing and wireless communication.

Mane V.A received the bachelor degree in electronics engineering also he was receive the Master of engineering in Electronics engineering. He is working as a assistant professor in Annasaheb Dange college of engineering, Ashta. His area of specialization in digital signal processing and embedded system.

DOI 10.17148/IJARCCE.2016.5617