International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 05 | May-2016 www.irjet.net p-ISSN: 2395-0072

Separating Voiced Segments from Music File using MFCC, ZCR and GMM

Mr. Prashant P. Zirmite[1], Mr. Mahesh K. Patil[2], Mr. Santosh P. Salgar[3],Mr. Veeresh M. Metigoudar[4]

[1,2,3,4]*Assistant Professor, Dept. of Electronics Engineering, DKETS TEI, Ichalkaranji, Maharashtra, India.*

Abstract -*Detection of frames that contains voice from the complete music file is key task in singing voice separation system.Music file is essentially a mixture of vocal and accompanied background instrument. To separate frames containing singing voice, the original file is framed and each frame is analyzed to find its voice content. Spectral features such as Mel-frequency Cepstral Coefficients (MFCCs), Log Frequency Power Coefficients (LFPC), Linear Prediction Coefficients (LPC) and temporal features as Zero Crossing Rate (ZCR), Pitch, timbre etc., can be used to detect voiced frames. In this paper MFCC and ZCR features are used to locate voiced frame. Gaussian Mixture Model (GMM) classifier is used to identify the voiced frames. Paper includes discussion and comparison of results by independent analysis of individual feature and their combination. Results indicate that higher classification accuracy is achieved for combination of features.*

Key Words:**GMM, MFCC, Music, Singing voice, ZCR**.

1.INTRODUCTION

There are some applications like: Singer Identification, Music Information Retrieval Systems (MIR), Lyrics to audio alignment, Music Structure Detection (Intro, Verse, Chorus, Bridge, Instrumental and Ending), identify singer characteristics, karaoke systems, identify language etc. in which separated singing voice is essential [1] [2] [3] [4].Music is an art whose medium is sound and silence. Here sound is singing voice with or without musical accompaniment, or only instrumental music. Shankar Vembu [5] et al. presented an approach using MFCC, LPFC and LPC to identify vocal frames in a music signal and designed a classifier to separatevocal–nonvocal frames. It can be seen from the results of their system that combinations of features give better performance when compared to individual features and the best performance of 77.24% efficiency resulted when multiple features were used as inputs to the network. Ziyou Xiong [6] et al. has compared six methods for classification of sports audio. With the use of features like: MPEG-7 audio features and Mel-scale Frequency Cepstrum Coefficients (MFCC). Maximum Likelihood Hidden Markov Models (ML-HMM) and Entropic Prior HMM (EP-HMM) were the classifiers they practiced. The improvement of about 8% is achieved by these combinations. As system for singing voice separation is again a case of speech separation, in a speech separation technique, Bachu R.G. [7] et al. explained how ZCR is useful in separating voiced and unvoiced part in speech. Their results suggested that zero crossing rates are low for voiced part and high for unvoiced part. In [8] L. R. Rabiner and M. R. Sambur deals with the problem of detecting speech in presence of noise (i.e. end point detection) using ZCR and energy. For different samples of 10 speakers no gross errors were detected in locating the end point. Ling Feng [9] et al. Provided systematic study of the performance of a set of classifiers, including Linear Regression (LR), Generalized Linear Model (GLM), Gaussian Mixture Model (GMM), reduced Kernel Orthonormalized Partial Least Squares (rKOPLS) and K-means by cross-validating both training and test setup.With experiments they proved that the classification performance depends heavily on the data sets, and the error rate varies between 13.8% and 23.9%

In this paper we implement a method to separate voiced frames from music file using MFCC as well as ZCR [10] as features and GMM as classifier. Input music files are divided in frames of length 25ms with overlap of 15ms. Number of samples in each frame depends on sampling frequency (Fs), i. e. if Fs is 16000 then 400 samples will be there in each frameof length 25ms. MFCC and ZCRare computed for each of above frames.Thenthese frames are classified as voiced or non-voiced using GMM [11] [12]. The paper is organised as follows. An overview of singing voice separation is given in section I. Section II describes features (MFCC and ZCR) and Classifier (GMM). Section III discusses database preparation, section IV includesresults. In section V parameter estimation is given and section VI provides Conclusion.

2.SYSTEM DESCRIPATION

2.1 Acoustic Features

MFCC is most commonly used acoustic feature in singing voice separation system. As human ear perceives sound logarithmically, and MFCC approximates the system response more closely to human audio system. Therefore MFCC is suitable tool for analysis of singing voice. The reason behind choice of ZCR is its ability to find utterances of voice in input signaland also for end point detection.

1. Mel-frequency Cepstral Coefficients (MFCCs)

In MFCC the frequency bands are positioned logarithmically. Considering constant changing nature of music signal MFCC is computed by dividing the input into number of overlapped frames.Assuming that on short time scale signal doesn`t change much, for each frame a MFCC vector is computed.

The steps involved in calculation of MFCC are shown in Fig.1.:

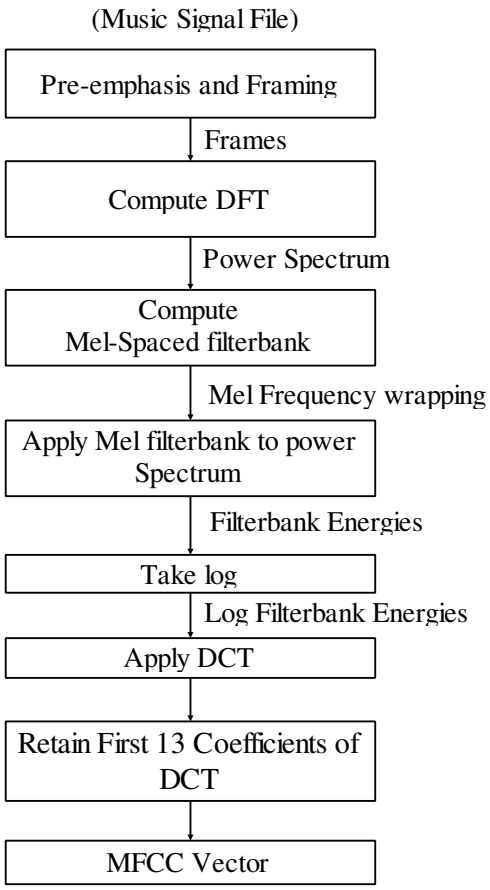

Fig. -1: Extraction of MFCC vector from music signal.

At pre-emphasis stage low frequency noise is reduced by passing signal through low pass filter. The signal is then framed in overlapping frames of 20ms to 30ms with 10ms to 15ms overlap [13]. After this power spectra of signal is calculated by taking DFT of each frame. In next step frames are passed through the Mel spaced filterbank to get filter bank energies. The width of triangular filters varies as per Mel frequency warping. Then the log total energy in critical band is included around center frequency of triangular filter. At the end the Discrete Cosine Transformer is used to calculate cepstral coefficients.

Mel frequencies are calculated from normal frequencies by using Equation (1)

$$Mel(f) = 1125 * ln(1 + \frac{f}{700}) \quad (1)$$

2. Zero Crossing Rate (ZCR)

Zero-crossing rate is one of the temporal feature which is used to identify voice contained frames in the process of singing voice separation. ZCR of an audio signal is a measure of the number of times the signal crosses the zero amplitude line by transition from a positive to negative or vice versa [7].The implementation of ZCR includes dividing the audio signal into temporal segments by using the window functionand zero crossing rate of each segment is computed using the relation as given below.

$$z = \sum_{i=1}^{w} \frac{|sgn(x_i) - sgn(x_i - 1)|}{2N} \quad (2)$$

Where $sgn\ (x_i)$ (signum function) indicates the sign of the i^{th} sample x_i and can have three possible values: +1, 0, -1 depending on whether the sample is positive, zeroor negative. The value for each window is collected to generate the ZCR feature vector having $W = \frac{n}{w}$ elements where *w* is window width and *n* is vector size.

$$Z = \cup_{j=1}^{w} z_j \quad (3)$$

2.2 Classifier

Classifiers role is to assign the frames of input data represented by their features to different categories as voiced or unvoiced. The proposed work emphasizes on choosing the Gaussian mixture model (GMM) as classifier [11].

The Gaussian mixture density is a weighted sum of M component densities, as follows:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (4)$$

Where$\vec{x}$is a *D*-dimensional random vector,$b_i(\vec{x})i$= 1, ...,*M*, are the component densities, and p_i are the mixture weights. Each component density is a Gaussian function.

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'\sum_i^{-1}(\vec{x} - \vec{\mu}_i)\right\} \quad (5)$$

$$\lambda = \{p_i, \bar{\mu}_i, \Sigma_i\} \quad (6)$$

With mean vector$\vec{\mu}_i$and covariance matrix$\sum_i$. The mixtureweights have to satisfy the constraint$\sum_{i=1}^{M} p_i = 1$. λ represents GMM model. The Expectation Maximization (EM) algorithm is used to estimate the parameters of the GMM. Resulted parameters of the GMM (including values of the mean vectors, covariance matrixes and weights) from the training procedures are then used to classify voiced frames so as to separate singing voice.

3. DATABASE PREPARATION

To perform experiment dataset is created. For this samples are taken from MIR-1k dataset [14] [15] [16] designed for research of singing voice separation.Three kinds of datasets are developed from the samplesof MIR-1k, as only voice, only music and mix which indicates a mixture of voice and music. Only voice is having pure vocal section of the song, similarly only music has pure music signal with no vocals, whereas the mixture is created by mixing voice and music at zero signal to noise ratio. Signals are having sampling frequency of 16000 samples per second. Different samples are chosen such as male voice, female voice, different musical accompaniments and singer with chorus.

4. IMPLEMENTATION AND RESULTS

4.1 Computation of MFCC Vector

The input signal is first framed into frames of 25ms with 10ms shift. Mean values of MFCC vector is calculated for all frames. As sampling frequency is 16000, each frame contains 400 samples, next frame is shifted by 160 samples. For each frame MFCC is calculated, result of MFCC is vector of order m by n, where m is the order of MFCC and n is number of frames. The spectrogramsfor all types of the input music signal is shown in Fig. 2.

As seen in Fig. 2 the region where the voice is present is having more spectral energy than the only music region.

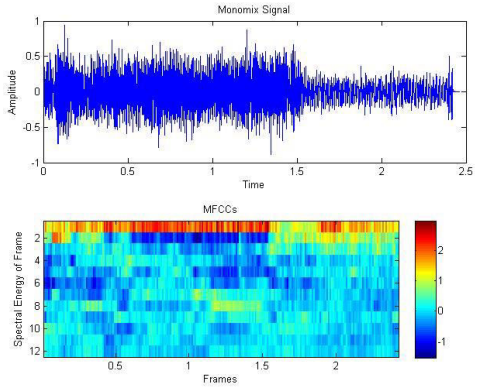

Fig.- 2: MFCCs for Mix signal

4.2 Computation of Zero Crossing Rate

Similar to MFCC, ZCR computation also first involves framing of the signal. Each frame is analysed for its zero crossing and ZCR is calculated for all the frames using Equation 2. Fig.3 shows ZCR values for mix signal. It is observed that the ZCR values ofvoice frames are more than music only frames.

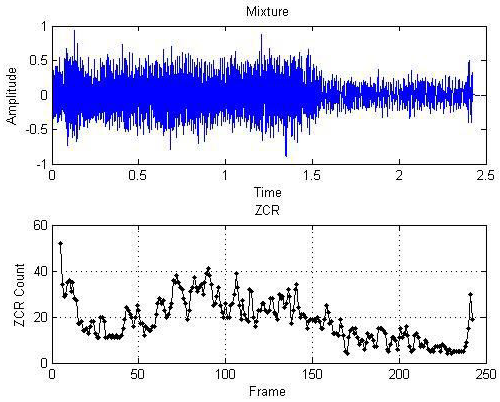

Fig.- 3: ZCR values for Mix type input.

4.3 GMM based classifier

The expectation maximization (EM) algorithmis used to estimate GMM parameters [9]. Feature vector generated at earlier stage acts as parametric distribution p(x) for GMM. The EM algorithm is an iterative process in which initialization of parameters is important. A prior model of initial values for GMM parameters like mean, variance and mixture weights are used at beginning. Then new model is calculated using Equation 7.

$$p(X|\bar{\lambda}) \geq p(X|\lambda) \qquad (7)$$

The above steps are iterated tillconvergence is reached.

A. Training

GMM is trained to identify voiced frames by using samples from training database. Training database consist of only voice samples of male and female singers. For these training samples MFCC and ZCR vectors are computed. GMM is separately trained for only MFCC and combine vector of MFCC and ZCR.The vocal threshold value is decided from trained value of GMM parameters. Here minimum value of mean vector is taken as threshold to identify voiced frame.

B. Testing

Input test signal is provided to system and GMM parameters are calculated. Mean vector is then compared with the threshold from training stage. If value for mean of frame is more than threshold, the frame is labeled as voice frame as shown in Fig.4. These voice frames are then separated as shown in Fig.5.

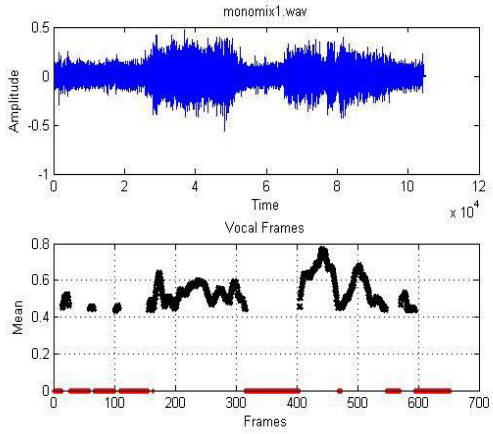

Fig. - 4: Frames labeled as voiced.

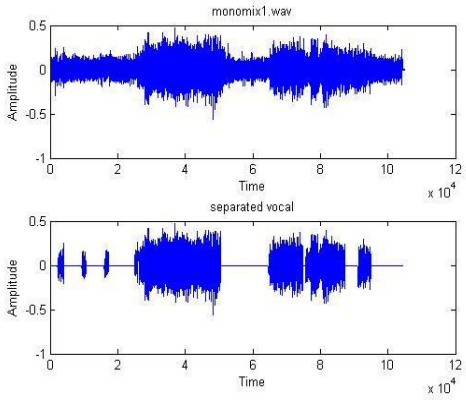

Fig. -5: Separated vocal frames.

5. PERFORMANCE ESTIMATION

To evaluate the system, proposed method is applied on ten input music signalsand the parameters like error rate and Signal to Noise Ratio (SNR) of system are calculated. Accuracy of the system is measuredfrom error rate values.Results of SNR before separation and after separation highlights the improvement because of combining multiple features.

5.1 Error Rate

Error rate is the measure to find how correctly the frames are classified. Here error rate is calculated by using Equation 8, as below

$$Error\ rate\ (\%) = (1 - Precision) \times 100 \quad (8)$$

Where,

$$Precision = \frac{Frames\ correctly\ labeled\ as\ Voiced}{Number\ of\ voiced\ frames\ provided} \quad (9)$$

Fig.6 gives error rate for different input music signals for MFCC, ZCR and their combination.

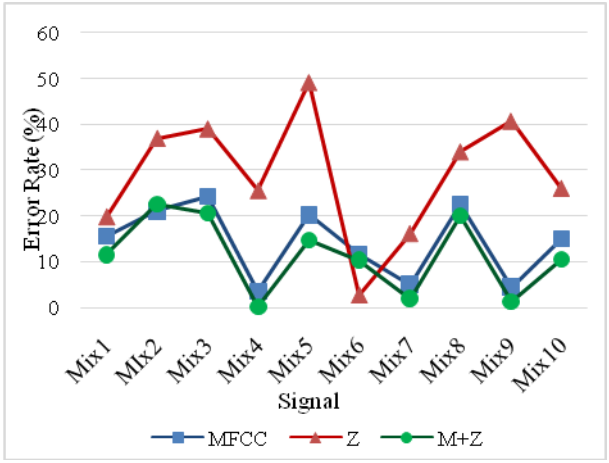

Chart-1: Error rate for ten sample of all three type.

Chart 1 clearly illustrates the error rate is minimum for combined features (MFCC and ZCR) than that of individual features.

5.2 Signalto Noise Ratio (SNR)

To quantify the performance of the system, we calculate the SNR before and after the separationusing Equation 10

$$SNR = 10log_{10}\left[\frac{\sum_n I^2(n)}{\sum_n (I(n)-O(n))^2}\right] \quad (10)$$

Where *I(n)* is the clean reference singing voice signal prior to mixing. In calculating the SNR after separation, *O(n)* is the output of the separation system. In calculating the SNR before separation, *O(n)* is the mixture signal.

Table 1 :SNR gain of voiced part from separated vocal and that of from input mixture signal

Signal	SNR (dB) Gain of Output			SNR of Input
	MFCC	*ZCR*	*MFCC+ZCR*	
Mix1	3.7813	4.5102	3.4749	2.7656
Mix2	6.0136	5.8745	5.7164	5.3487
Mix3	5.4084	5.4135	5.2441	4.9911
Mix4	5.1706	7.6809	4.9584	4.4663
Mix5	5.2568	13.8495	5.0721	4.8097
Mix6	4.7374	7.4922	4.6321	4.3831
Mix7	5.2185	4.794	4.9598	4.5606
Mix8	5.1889	5.3659	5.0485	4.8234
Mix9	5.3429	11.8026	5.1784	4.8358
Mix10	4.4447	5.3333	3.842	3.0173

The computed SNR gain (in dB) for voiced part of signal after separation together with the SNR gain of the applied input signal is presented in Table 1. SNR is calculated separately for MFCC, ZCR and their combination. Table I. clearly indicates that the SNR gain of the system output nearly equals to that of input, when we combine both the features.

In Table II classification accuracy of implemented method is given. When using only ZCR accuracy is comparatively

less for nearly all input signals. Use of MFCC provides moderate accuracy. But higher accuracy is observed when we combine both features.

Table -2:classification accuracy of separate voiced frames

Signal	Classification Accuracy (%)		
	MFCC	*ZCR*	*MFCC+ZCR*
Mix1	84.38	80.11	88.41
Mix2	78.86	62.94	77.36
Mix3	75.74	60.89	79.21
Mix4	96.4	74.41	99.83
Mix5	79.74	50.75	85.29
Mix6	88.31	97.24	89.58
Mix7	94.9	83.77	98.06
Mix8	77.42	65.91	79.84
Mix9	95.51	59.2	98.8
Mix10	84.95	73.89	89.51
Average	85.62	70.91	88.59

6. CONCLUSION

This paper reveals that to detect and separate voiced frames from music file MFCCs and their derivate are the much appropriate. Otherfeatures describing the temporal characteristics e.g. ZCR can also be used.

We achieved average classification accuracy of 85.62% when only MFCC is used as feature, 70.91% with ZCR individually, and for combined features (MFCC + ZCR) an accuracy of 88.59% is achieved in classifying the frames. Clearly higher efficiency is achieved for combined approach of both temporal and spectral features for the estimated task

REFERENCES

[1] H. Fujihara and M. Goto, "Layrics-to-Audio aligenment and its application," Dagstuhl Follow Ups, vol. 3, pp. 23-36, 2012.

[2] N. C. M. M. S. K. Changsheng Xu, "Automatic Structure Detection for Popular Music," IEEE MultiMedia,, vol. 13, no. 1, pp. 66-71, 2006.

[3] M. Khadkevich, "Music Signal Processing For Automatic Extraction Of Harmonic And Rhythmic Information," DISI- University Of Trento, Trento, December 2011.

[4] Michael J. Henry, "Learning Techniques for Identifying Vocal Regions in Music Using the Wavelet Transformation, Version 1.0," Washington, 2011.

[5] S. Baumann and S. Vembu, "Separation Of Vocals From Polyphonic Audio Recordings," in International Symposium Music Information Retrieval(ISMIR`05), London, 2005.

[6] R. Radhakrishnan, A. S. Divakaran and Z. Xiong, "Comparing MFCC And Mpeg-7 Audio Features For Feature Extraction,Maximum Likelihood Hmm And Entropic Prior Hmm For Sports Audio Classification," in roc. IEEE Int. Conf. Multimedia Expo, 2003.

[7] K. S. A. B. B. B. Bachu R.G., "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal," Journal of Acoustic Society of Electrical Engineering, 2008.

[8] L. R. R. a. M. R. Sambur, "An Algorithm for Determining the Endpoint of Isolated Utterances," The Bell Systems Technical Journal, vol. 54, no. 2, pp. 297-315, February 1975.

[9] A. Nielsen, L. Kai and H. L. Feng, "Vocal Segment Classification In Popular Music," in International Symposium Music Information Retrieval(ISMIR`08)(Timbre), 2008.

[10] L. S. K. B. Arif Ullah Khan, "Hindi Speaking Person Identification Using Zero Crossing Rate," International Journal of Soft Computing and Engineering (IJSCE), vol. 2, no. 3, July 2012.

[11] C. L. Hsu, D. Wang, J. S. Roger Jang and K. Hu, "Tandem algorithm for Singing Pitch extraction and voice separaation from music accompaniment," IEEE Transaction on Audio, Speech and Languuage processing, vol. 21, no. 1, pp. 1482-1491, January 2013.

[12] P. P. R. G. F. B. Alexey Ozerov, "One Microphone Singing Voice Separation using Source-Adapted Models," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 16-19 October 2005.

[13] J. Lyons, " practicalcryptography," 2009-2012. [Online]. Available: http://practicalcryptography.com/.

[14] J.-S. R. J. Chao-Ling Hsu, "http://www.mirlab.org," [Online]. Available: http://mirlab.org/dataset/public/MIR-1K.rar. [Accessed 07 09 2013].

[15] C.-L. Hsu and J.-S. R. Jang, "Dataset for singing voice separation," [Online]. Available: http://mirlab.org/dataset/public/MIR-1K.rar.

[16] S. Chen, S. Paris, M. Hasegawa and P.-S. Huang, "Singing-Voice Separation From Monaural Recordings Using Robust Principal Component Analysis," IEEE ICASSP, Vols. 987-1-4673-0046-9, no. 12, pp. 57-60, 2012.