Signal Processing for Segmentation of Vocal and Non-Vocal Regions in Songs: A Review

[1]A Bonjyotsna & [2]M Bhuyan
Deptt. of Electronics and Communication Engineering
Tezpur University
Tezpur-784028, India
[1] *ananyab@tezu.ernet.in*
[2] *manab@tezu.ernet.in*

Abstract- **It has been of great importance to us to identify a singer in a song or from collection of songs. Speaker identification is being widely used in organizing, browsing and retrieving music collections, and its speech processing techniques are well established. But the inability to delete the background accompaniment fully has set a limitation in extracting singing voice features appropriately. Singing voice identification is basically done on the basis of three techniques viz. locating vocal/non-vocal segment from the song, feature extraction of the vocal segment and statistical classification. This paper attempts to review on the various techniques developed for detection of vocal/non-vocal segments and tries to find out scope of improvement to formulate a generalized vocal detection technique.**

I. INTRODUCTION

Singing voice of human is the oldest musical instrument that has been known so far. In the music world, the singing voice plays a very important role. It is also the most complex instrument from the acoustic point of view. This is due to the rapid acoustic variation involved in the singing process. No other instrument exhibits the amount of physical variation of the human voice. Speech signals are composed of a sequence of sounds and the sequence of sounds are produced as a result of acoustical excitation of the vocal tract when air is expelled from the lungs. VOICED speech is produced when the vocal cords play an active role (i.e. vibrate) in the production of a sound (eg: /a/,/e/,/i/). UNVOICED speech is produced when vocal cords are inactive (eg: /s/,/f/). The Vocal Tract is a non-uniform acoustic tube that is terminated at one end by the vocal cords and at the other end by the lips. The cross-sectional area of the vocal tract determined by the positions of the tongue, lips, jaw and velum depends on lips, tongue, jaw and velum [1].

There are also mixed sounds that use both voiced and unvoiced to create the overall sound. Since speech processing techniques are well established, it can be therefore applied to singing voice signal although there are prominent differences in both the signals. Both speech and singing are comprised of voiced (e.g. vowels) and unvoiced (e.g. stops and fricatives) utterances and voiced utterances usually exhibit predominantly harmonic structure in the frequency domain. The time ratio for voiced/unvoiced/silent phonation in speech is approximately 60%/25%/15% as compared with the nearly continuous 95% for singing [2], [3]. However, the intensity and dynamic range of the singing voice is usually greater than speech.

The most obvious difference between singing and speech is the variation of pitch. The pitch of normal speech ranges from 80 and 400 Hz, while that of singing can be from 80 to 1000 Hz [2]. In singing, pitch may be further modulated using a frequency near 4–8 Hz, which results in a phenomenon called *vibrato* [4], [6]. Vibrato occurs rarely in speech, but it is still present in very few spoken utterances [5]. The formant structures between speech and singing are also quite different.

Rosenau analyzed synchronous recordings of acoustic and electroglottographic signals of several singers [7]. He found that a significant difference between singing and speech lies in the vowel quality for the second and third formants. It is observed that there is a tendency for singer, especially in operatic singing, to group the third, fourth and fifth formants together, so as to enhance the audibility of the singing voice in the presence of instruments [7]. Also, singing voice synthesis has been investigated in several studies. Some of them develop methods for converting speech into singing [8]. Classically-trained singers are taught that vowel sounds should be sustained for long between consonants since they are the most efficient and audible sounds (consisting of several resonant peaks), which is especially important for being heard over other instruments.

II. SINGER IDENTIFICATION

Sensitivity to the human voice reception has evolved as an improvement of our auditory physiology and perceptual apparatus. Once we hear to the sound of a person's speaking voice, it is relatively easier to identify that voice almost without any training. The same happens with a singing voice. Once we become familiar with the sound of a particular singer's voice, we can usually identify the voice, even when hearing a piece for the first time.

Generally songs are discriminated based on singer's voice since one could subconsciously identify the singer when he hears a song. Therefore, all Karaoke systems usually categorize music by the names of singers, and so as the music stores. As the singing voice is important, the representation of its characteristics is necessary for content-based Music Information Retrieval (MIR). Consequently, singer identification is one of the most important tasks in content-

978-1-4673-4862-1/13/$31.00 ©2013 IEEE

based fields. However, till now, the MIR system is based on text tags of singer's names and song titles, not on characteristics of his or her voice. If the singer's name is not known, a user could hardly find songs he wants. Hence, study based on using vocal segment in a song for retrieval is rather necessary. In other words, singer identification is to retrieve the names of singers through the singers' voice.

Singer identification can be mainly divided into three parts:

1) Locating vocal/non-vocal segment from the song,
2) Feature extraction of the vocal segment and
3) Classification

III. Review of Research on detection of vocal/nonvocal segments

Before identifying the singer it is necessary to locate the singer's voice in a song. Maximum of the songs start with a piece of instrumental accompaniment known as prelude in musical terms after which the singing voice comes into play. Therefore, it is necessary to detect the vocal region in the song in order to extract the singer's voice characteristics and to avoid the non-vocal region which includes the instrumental accompaniment. The vocal region of a specified length is collected for feature extraction. The vocal region also includes the instrumental music which sometimes predominates the region by masking which means energy of one frequency band will mask lesser energies in the adjacent frequency bands. Hence, it is also required to minimize the effect of accompaniment in the vocal region for extracting more robust features of the singing voice. There have been many researches going on to accurately detect the vocal region and also to improve the reduction of background music.

Many different approaches have been made by different researchers to classify the vocal and nonvocal parts. Basically most of them have applied statistical classifier by modeling the vocal and nonvocal part separately. For classification, two main components are required which are (1) features and (2) classifiers. Different features have been explored for singing voice detection. These features are Mel-frequency Cepstral Coefficients (MFCC), Linear Predictive Coefficients (LPCs), Perceptual Linear Prediction Coefficients (PLPs) and the Harmonic Coefficients. MFCC, LPC, and PLP are also widely used for general sound classification tasks and they are called short term features because they are calculated in short time frames. Similarly different classifiers have also been explored including Gaussian Mixture Models (GMM), Hidden Markov Model (HMM), Support Vector Machines (SVMs) and multilayer perceptrons (MLPs). Some researchers have also applied the direct energy distribution criteria and filtering to detect the vocal segment.

In 2001, Berenzweig and Ellis [9] used a speech recognizer's classifier to distinguish vocal segments from accompaniment. They have used a neural network acoustic model, trained to discriminate between context-independent phone classes of natural English speech, to generate a vector of posterior probability features (PPFs) which were used as the basis for further calculations. When compared to non-vocal accompaniment within music, the PPFs will tend to shift more towards the singing voice due to its resemblance with the natural speech. The basis of the segmentation scheme of this work is to detect this characteristic shift. Three broad feature sets for this detection are considered: (1) direct modeling of the basic PPF features, or selected class posteriors; (2) modeling of derived statistics, such as classifier entropy, that should emphasize the differences in behavior of vocal and instrumental sound; and (3) averages of these values, exploiting the fact that the timescale of change in singing activity is rather longer than the phonetic changes that the PPFs were originally intended to reveal, and thus the noise robustness afforded by some smoothing along the time axis can be usefully applied. Their classifier achieved around 80% classification accuracy at the frame level. In [10] Berenzweig, Ellis and Lawrence use a neural network trained on radio recordings to similarly segment songs into vocal and non-vocal regions. By focusing on voice regions alone, they were able to improve artist identification by 15%.

Although Berenzweig and Ellis used speech recognizer to detect singing voice, due to their significant differences it may not be appropriate to use the techniques used for speech. Tsai and Wang [11] in 2006 proposed to construct a statistical classifier with parametric models trained using accompanied singing. This approach is based on the observation that there is a significant difference in spectral distribution between vocal and instrumental sound. Compared to the singing voices, the instrumental-only sounds have less salient harmonics and spread their energy more widely [12].

The vocal/nonvocal classifier unit shown in the Fig.1 consists of a front-end signal processor that converts digital waveforms into spectrum-based feature vectors, and a back-end statistical processor that performs modeling, matching and decision making. The feature vectors used here are Mel-scale frequency cepstral coefficients. (MFCCs), which are typically computed using a fixed length sliding window of 10 ms to 40 ms, also called a frame.

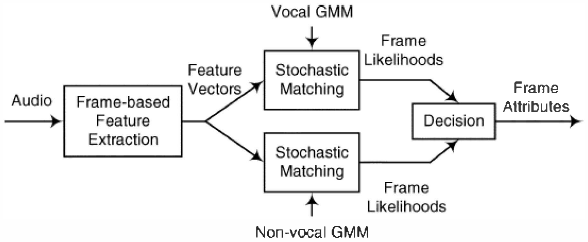

Fig. 1 Vocal/Non-vocal segmentation [11]

The back-end statistical processor operates in two phases: training, and testing. During training, a music database with manual vocal/non-vocal transcriptions is used to form two separate Gaussian mixture models (GMMs): a vocal GMM, and a non-vocal GMM. Each model consists of several mixture weights, mean vectors and covariance matrices. In the testing phase, the classifier takes as input the feature vectors extracted from an unknown recording, and produces as output the frame log-likelihoods for the vocal and non-vocal GMM, respectively [11].

Since majority of energy in the singing voice falls between 200Hz and 2000Hz, Kim and Whitman [13] has developed a straight forward method to detect energy within the frequencies bounded by the range of vocal energy. They have used a simple Chebychev infinite-impulse response (IIR) digital filter of order 12 to filter the audio signal bandpassing the vocal range to pass through while attenuating other frequency regions. The response of such a filter is shown in Fig.2.

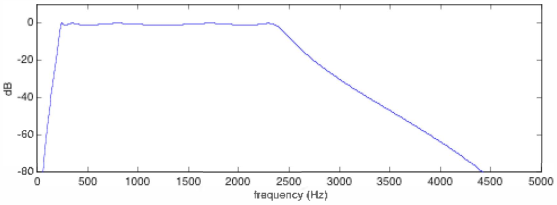

Fig. 2: Vocal enhancement filter frequency response [13]

This filter attenuates the instruments that fall outside of this frequency region, such as bass and cymbals. But in popular music, the drums also fall under the energy region of the singing voice. So another measure was taken to discriminate the voice from these other sources. Since it is known that singing voice is highly harmonic and other sources such as drums are not as harmonic as the singing voice, this difference could be exploited and an inverse comb filterbank was used to detect high amounts of harmonic energy. The block diagram and frequency response of a simple inverse comb filter is shown in Fig.3.

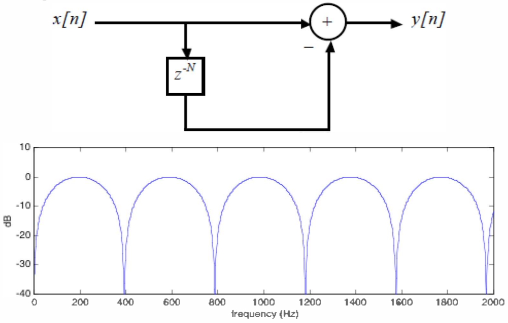

Fig. 3 Block diagram of the simple inverse comb filter (top). The frequency response of an inverse comb filter, tuned to ~400 Hz (bottom). The spacing of the attenuated frequencies is determined by the delay parameter N [13].

By taking the ratio of the total signal energy to the maximally harmonically attenuated signal, harmonicity is measured (i.e. how harmonic the signal is within the analysis frame) and it is given by the equation (1) [13],

$$H = \frac{E_{original}}{\underset{i}{\underbrace{\min}}(E_{filtered,i})} \quad (1)$$

By thresholding the harmonicity against a fixed value, we have a detector for harmonic sounds which correspond to that of the singing. The results of their experiments performed on a set of 20 songs show that as the threshold is increased the detection becomes more accurate. The experimental results are shown below.

TABLE I
Performance of vocal detector at multiple harmonicity thresholds [13]

H Threshold	Vocal Segments	Non-vocal Segments	All Segments
2.0	55.4%	53.1%	55.4%
2.3	40.5%	69.2%	55.1%
2.6	30.7%	79.3%	54.9%

On the other hand Zhang [14] has proposed an effective method to identify the starting point of the singing voice in the song. After detecting the starting point a fixed length between 10 to 30seconds of audio signal is taken in the song from the starting point of the singing voice and used as testing data. To detect the starting point, Zhang used four kinds of audio features together-

1) *The Energy Function:* It represents the amplitude variation over the time of the audio signal. The start of singing voice (arrow in Fig 4) is normally reflected as a sudden rise in the energy level of the audio signal.

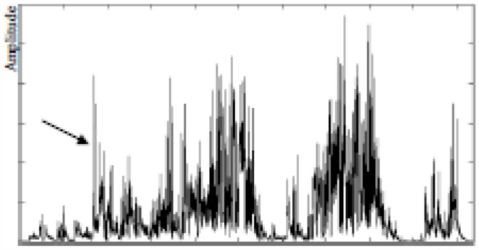

Fig. 4 Energy function of a song [14]

2) *The average Zero-crossing rate (ZCR)*: it is a simple measure of the frequency content of the audio signal [1]. While ZCR values of instrumental music are normally within a relatively small range, the singing voice is often indicated by high amplitude ZCR peaks (arrow in Fig 5) resulted from pronunciations of consonants.

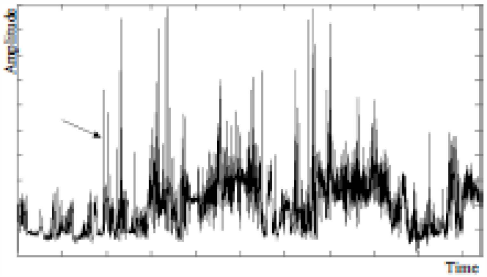

Fig. 5 Average Zero-crossing rates of a song [14]

3) *The Harmonic Coefficient*: The harmonic coefficient, as defined in [15] H_a is calculated by the maximum of

the average autocorrelation value in the time-domain and the frequency domain, which gives a good indication of existing harmonic components. Suppose $R_T(\tau)$ is the temporal autocorrelation for candidate pitch τ, $R_S(\tau)$ is the corresponding spectral auto-correlation, then to improve robustness, $R_T(\tau)$ and $R_S(\tau)$ are combined as:

$$R(\tau) = \beta \cdot R_T(\tau) + (1-\beta) \cdot R_S(\tau) \qquad (2)$$

where β =0.5 turned out to perform well. Then, H_a is defined as-

$$H_a = \max R(\tau) \qquad (3)$$

According to [15], the singing signal in general has higher values of the harmonic component, compared to the instrumental music. Therefore, the start of the singing voice may be indicated by an abrupt increase in the H_a value.

4) *The Spectral Flux*: It is defined as the 2-norm of the frame-to-frame spectral amplitude difference vector given by-

$$F_n = \left\| |X_n(\omega)| - |X_{n+1}(\omega)| \right\| \qquad (4)$$

where $|Xn(\omega)|$ is the magnitude spectrum of the nth frame of the audio signal. The start of singing voice is often indicated by the appearance of high peaks in the spectral flux value (arrow in Fig 6), because the voice signal tends to have higher rate of change than instrumental music.

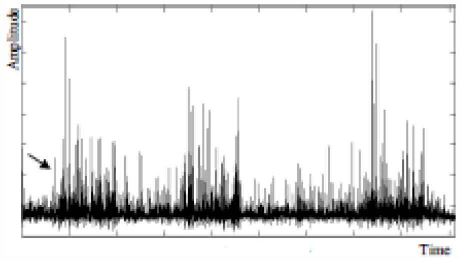

Fig. 6 Spectral flux of a song.[14]

With this procedure it was proved to be able to discard most of the prelude in songs and extract segments consisting of typical voice signals of the singer.

In 2007 Nwe and Li [16] put forwarded a novel approach to extract the vocal segments. Firstly the vocal and nonvocal parts are trained using HMM models by manually annotated songs and then performed classification between vocal and nonvocal segments. They have also used the adaptive classifier to adapt to the song-specific acoustic characteristics. In order to avoid inaccurate vocal detection they have also proposed a hypothesis test process to validate each of the vocal detection decisions. Vocal/ nonvocal classification decisions are made on every sub-segment of 1 s. The error rate is shown in the table below as the number of error detection over the total number of trials from a test set.

TABLE II
Error rates (E_R %) of vocal detection (N= Nonvocal, V= Vocal , Ave= Average) [16]

Vocal Detector model	N	V	Ave
MM-HMM	27.2	7.5	17.3
MM-HMM + Hypothesis test	7.7	9.7	8.7
Baseline	22.4	17.4	19.9

The hypothesis test rejects 3.0% of unsure segments. The hypothesis test greatly improves the performance of vocal detection, achieving a 49.7% relative error reduction (from 17.3% to 8.7%). Therefore, this work distinctly eliminates the problems regarding mismatch between normal speech and singing which were faced by Berenzweig *et al.*

In another work, Li and Wang [17] have proposed, HMM and rule-based post-processing since the vocal or nonvocal portions sustains a certain amount of time so that the short-term classification does not jump back and forth rapidly. They have also proposed a novel method for extracting vocal portions that takes into account the rhythmic aspects of music signals. To partition the input into useful portions, they have used a simple spectral change detector proposed by Duxbury *et al* [21] and then pool the likelihoods over all the frames of a portion and classify the portion into the class with larger overall likelihood. For comparing the results they refer to the energy ratio of singing voice to accompaniment as signal-to-noise ratio (SNR). Nwe *et al.* [18] pointed out that different sections of a song (intro, verse, chorus, bridge, and outro) have different SNRs, and a singing voice detector needs to handle different sections properly. To address this problem, they trained a classifier with samples mixed in different SNRs. In this way, the classifier is trained over a range of SNRs. They performed tenfold cross validation to access the overall performance of the proposed detection method. Fig. 7 depicts the documentation of the results of singing voice detection error rate.

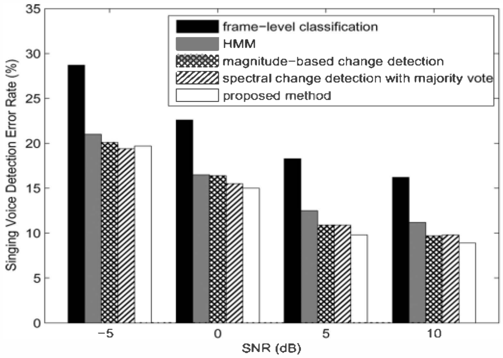

Fig.7 Performance of the proposed singing voice detection algorithm in different SNRs. For comparison, the performances of four alternative methods are also shown.[17]

Recently Tsai and Lin [19] has formulated techniques on how to remove the background music accompanied with the singer. In order to segment vocal and nonvocal region they have adapted the process of [11] which is described earlier. After segmentation their work aims to find the solo voice from the background accompaniment. They have achieved it by transforming the cepstrum of an accompanied singing into a solo voice one. The general idea that has been considered here is that a solo voice is superimposed by some accompaniments. They have generated databases by manually superimposing accompaniments on solo voices. By modeling the variations of voice cepstrum before and after background music is added in it, it is able to transform an accompanied voice to a solo one thereby making it more reliable for feature extraction.

IV. CONCLUSION

Although the vocal portions are efficiently and adaptively detected in most of the genres there still lies the problem in nullifying the accompaniment. The more effectively the voice is extracted, the more accurate it becomes for identification of the singer with very little influence of the background music. But the challenge is to remove the accompaniment part completely to get the vocal part exclusively. Most of the works have performed modeling of the instrumental region so as to classify with the vocal region and hence segregating the two. The inability to delete the background accompaniment fully has set a limitation in extracting singing voice features appropriately. There are some other singing styles viz. falsetto, growl and undertone singing which are still confusing. Such issues can be solved by intelligent approaches such as-reasoning, learning and knowledge [20]. Therefore in order to formulate a generalized vocal detection technique there are still more works need to be done for higher accuracy for singer identification.

REFERENCES

[1] L.R. Rabiner and R.W.Schafer, *Digital Processing of Speech Signals*, Prentics-Hall, Inc., New Jersey, 1978.

[2] P. Rao, “Musical information extraction from the singing voice,” in *Proc. National Conf. Signal Image Process.*, 2009.

[3] Tsai, W.-H.; Lee, H.-C.; , "Singer Identification Based on Spoken Data in Voice Characterization," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.20, no.8, pp.2291-2300, Oct. 2012 doi: 10.1109/TASL.2012.2201473

[4] D. Gerhard, “Pitch-based acoustic feature analysis for the discrimination of speech and monophonic singing,” *J. Canadian Acoust. Assoc.*,vol. 30, no. 3, pp. 152–153, 2002.

[5] D. Gerhard, “Computationally measurable differences between speechand song,” Ph.D. dissertation, Simon Fraser Univ., Burnaby, BC,Canada, 2003.

[6] Nwe, T. L.; Li, H.; , "Exploring Vibrato-Motivated Acoustic Features for Singer Identification," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.15, no.2, pp.519-530, Feb. 2007 doi: 10.1109/TASL.2006.876756

[7] S. Rosenau, “An analysis of phonetic differences between German singing and speaking voices,” in *Proc. 14th Int. Congr. Phon. Sci.(ICPhS)*, 1999.

[8] T. Saitou, M. Goto, M. Unoki, and M. Akagi, “Speech-to-singing synthesis: Vocal conversion from speaking voices to singing voices by controlling acoustic features unique to singing voices,” in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA’07)*, 2007, pp. 215–218.

[9] Berenzweig, A.L.; Ellis, D.P.W.; , "Locating singing voice segments within music signals," *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the* , vol., no., pp.119-122,2001 doi:10.1109/ASPAA.2001.969557

[10] A. Berenzweig, D. P. W. Ellis, and S. Lawrence, “Using voice segments to improve artist classification of music,” in *Proc. 22nd AES Int. Conf.*, Espoo, Finland, Jun. 2002.

[11] Wei-Ho Tsai; Hsin-Min Wang; , "Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.14, no.1, pp. 330- 341, Jan. 2006 doi:10.1109/TSA.2005.854091

[12] P. R. Cook, “Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing,” Ph.D., Stanford Univ., Stanford, CA, 1990.

[13] Y.E.Kim and B.Whitman,”Singer identification in popular music recordings using voice coding features,” in *Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR 2002),* 2002, pp. 164-169.

[14] Tong Zhang; , "Automatic singer identification," *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on* , vol.1, no., pp. I- 33-6 vol.1, 6-9 July 2003 doi: 10.1109/ICME.2003.1220847

[15] W. Chou and L. Gu,”Robust singing detection in speech/music discriminator design,” *Proc of ICASSP’01*,Vol 2, pp. 865-868, Salt Lake city, May 2001

[16] Nwe, T. L.; Li, H.; , "Exploring Vibrato-Motivated Acoustic Features for Singer Identification," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.15, no.2, pp.519-530, Feb. 2007 doi: 10.1109/TASL.2006.876756

[17] Yipeng Li; DeLiang Wang; , "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.15, no.4, pp.1475-1487,May2007 doi: 10.1109/TASL.2006.889789

[18] T. L. Nwe, A. Shenoy, and Y. Wang, “Singing voice detection in popular music,” in *Proc. 12th Annu. ACM Int. Conf. Multimedia*, 2004, pp. 324–327.

[19] Wei-Ho Tsai; Hao-Ping Lin; , "Background Music Removal Based on Cepstrum Transformation for Popular Singer Identification," *Audio, Speech, and Language Processing, IEEE Transactions on* , vol.19, no.5, pp.1196-1205,July2011 doi: 10.1109/TASL.2010.2087752

[20] M Bhuyan, “Intelligent Instrumentation: Principles and Applications”, Taylor and Francis, USA,,pp.445-482, ISBN: 978142008953, 2010

[21] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, “Complex domain onset detection for musical signals,” in *Proc. 6th Conf. Digital Audio Effect (DAFx-03)*, London, U.K., 2003.