ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Audio Signal Separation and Classification: A Review Paper

Silk Smita[1], Sharmila Biswas[2], Sandeep Singh Solanki[3]

Birla Institute of Technology, Mesra, Ranchi, India[1,2,3]

ABSTRACT: Music signals are not solely characterized because of other mixed audio signals. Mixed audio signals contain music signals mixed with speech signals, voice and even background noise.Thus, mixed signals need to classify separately. Researchers have developed many algorithms to solve this problem keeping in mind with their characteristic features of music signals: by timbre, harmony, pitch, loudness etc. The algorithm ICA (Independent component analysis) uses basis of Blind source separation, HSS(Harmonic structure stability), "SOSM" APPROACH (Second Order Statistical Measures Approach), Sinusoidal Parameters based audio classification using FDMSM etc. are some of the mixed signal classification algorithms. This paper highlights all these existing methods and their experimental results.

KEYWORDS: Harmonic structure stability;ICA; HSS; FDMSM; SOSM; audio separation; voice; music

I. INTRODUCTION

In this digital world, efficient management of the digital content of audio data has become an important area for the researcher. However, due to time-consuming and expensive indexing and labeling that are performed manually, an automatic content-based classification system of mixed type audio signal is a must for various applications of signals used these days. Separation of voice (speech), music and noise in a mixed audio signal is an important problem in music research. Here, *voice* means the singing voice in a song, *music* means the instrument. Separation of mixed signal is helpful for many other music researches, such as Music Retrieval, Classification and Segmentation, Multi-pitch estimation, etc. Along with audio signal separation, audio classification is important due to following reasons:

(a) Different audio types should be processed differently.

(b) The searching space after classification is reduced into a particular subclass during the retrieval process.

This approach reduces the computational complexity and one can choose specified field of research for that particular signal. For example, speech signals can be used for speech recognition, music signals for music analysis and voice for speaker recognition.

II. RELATED WORK

In this review, music signals are of prime interest. Music signals are characterized according to their pitch, loudness, duration, harmonic structure and timbre. Timbre is mainly used for the recognition of sound sources. Music signals are more “ordered” than voice. The entropy of music is much constant in time than that of speech [1].

Among various methods [1-16] for signal separation, harmonic structure model [4,5] has extended multi-pitch estimation algorithm used to extract harmonic structures, and clustering algorithm is used to calculate harmonic structure models. Then, signals are separated by using these models to distinguish harmonic structures of different signals.

ICA (Independent Component Analysis) is one of the methods for extracting individual signals from mixtures. Its power resides in the physical assumptions that the different physical processes generate unrelated signals. The simple and generic nature of this assumption allows ICA to be successfully applied in the diverse range of research fields [5]. Vanroose used ICA to remove music background from speech by subtracting ICA components with the lowest entropy [6]. General signal separation methods do not sufficiently utilize the special character of music signals. Gil-Jin and Te-

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Won proposed a probabilistic approach to single channel blind signal separation [7], which is based on exploiting the inherent time structure of sound sources by learning a priori sets of basis filters. In Harmonic Structure Stability approach [4], training sets are not required, and all information is directly learned from the mixed audio signal. Feng et al. applied FastICA to extract singing and accompaniment from a mixture [11].

Based on different features and characteristics, audio classification is used to classify most audio into speech, music and noise [2-3]. Many features are used including Cepstrum [4], [5], power spectrum or ZCR proposed in [6] using Support Vector Machine (SVM) classifier. Sinusoidal parameter based audio classification method [5] introduce a new unsupervised Feature selection method; in order to determine which features are optimum is sense of high accuracy in classification. Using only these obtained features it is demonstrated that acceptable classification accuracy is achievable. In addition, the SVM and Relevance Vector Machine (RVM) are employed since they were already proven to be the best classifiers for musical genre recognition in many literatures. This classification system categorizes audio based on new audio features, sinusoidal parameters denoting harmonicity as well as Subband energy. These sinusoidal parameters jointly represent harmonicity ratio (HR) and energies of different frequency subbands.

Another method used for classification is Statistical Discrimination and Identification of Some Acoustic Sounds [6].This method finds a pre-processing technique which allows discriminating between the different sounds. This technique is based on the similarity measure μ_{Gc} introduced by Bimbot and used for speaker recognition tasks.

III. METHODOLOGY

A. INDEPENDENT COMPONENT ANALYSIS[9]

Blind Signal Separation (BSS) or Independent Component Analysis (ICA) is the identification & separation of mixtures of sources with little prior information. Non-gaussianity is used to estimate the ICA model. ICA involves two basic steps

(a) "Nonlinear decorrelation"- In this step, matrix ***W is calculated*** so that for any $i \neq j$, the components $y_i$and $y_j$are uncorrelated, *and* the transformed components $g(y_i)$ and $h(y_j)$ are uncorrelated, where g and h are some suitable nonlinear functions."

(b) "Maximum non-gaussianity"- The local maxima of non-gaussianity of a linear combination $\boldsymbol{y=Wx}$under the constraint that the variance of $\boldsymbol{x}$ is constant. Each local maximum gives one independent component.

B. HARMONIC STRUCTURE STABILITY ANALYSIS[4,5]

This algorithm finds the average harmonic structure of the music, and then separate signals by using it to distinguish voice and music harmonic structures. Steps involved in this algorithm are preprocessing, harmonic structure extraction, music Average Harmonic Structure analysis, separation of signals.

S(t) is a monophonic music signal.

$$s(t) = \sum_{r=1}^{R} A_r(t)\cos[\theta_r(t)] + e(t)$$

$$A_r(t) \text{ and } \theta_r(t) = \int_0^t 2\pi r f_o(\tau)\, d\tau$$

are the instantaneous amplitude and phase of the r^{th}harmonic, respectively, R is the maximal harmonic number, $f_0(\tau)$ is the fundamentalfrequency, e(t) is the non-harmonic or noise component.

Divide s(t) into overlapped frames and calculate f_0^l and $A_r^l$by detecting peaks in the magnitude spectrum. A_r^l=0, if there doesn't exist the r^{th} harmonic. l = 1, . . . , L is the frame index. f_0^l and $[A_1^l, \ldots, A_R^l]$ describe the position and amplitudes of harmonics. Normalize A_r^l by multiplying a factor $\rho^l = C/A_1^l$(C is an arbitrary constant) to eliminate the influence of the amplitude. Translate the amplitudes into a log scale, because the human ear has a roughly logarithmic sensitivity to signal intensity. Harmonic Structure Coefficient is then defined as an equation. The timbre of a sound is mostly controlled by the number of harmonics and the ratio of their amplitudes, so $B^l = [B^l_1, \ldots, B^l_R]$, which is free from the fundamental frequency and amplitude, exactly represents the timbre of a sound. In this paper, these coefficients are used to represent the harmonic structure of a sound. Average Harmonic Structure and Harmonic

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Structure Stability are defined as follows to model music signals and measure the stability of harmonic structures.

Harmonic Structure B^l, B^l_i is Harmonic Structure Coefficient:

$B^l = [B^l_1, \ldots, B^l_R]$, $B^l_i = \log_{(\rho^l A_i^l)} / \log_{(\rho^l A_1^l)}, i=1,\ldots,R$

Average Harmonic Structure (AHS):

$$\overline{B} = \frac{1}{L}\sum_{l=1}^{L} B^l$$

Harmonic Structure Stability (HSS):

$$HSS = \frac{1}{R}\frac{1}{L}\sum_{l=1}^{L}\left\|B^l - \overline{B}\right\|^2 = \frac{1}{RL}\sum_{r=1}^{R}\sum_{l=1}^{L}(B_r^l - \overline{B}_r)^2$$

AHS and HSS are the mean and variance of B^l. Since timbres of most instruments are stable, B^l varies little in different frames in a music signal and AHS is a good model to represent music signals. On the contrary B^l, B^l varies much in a voice signal and the corresponding HSS is much bigger than that of the music signal.

C. Second Order Statistical Measures Approach[7]

This method [5], based on mono-Gaussian model [3], uses some measures of similarity, which are called Second Order Statistical Measures (SOSM). These measures are used in order to recognize the speaker at each segment of the speech signal.

Two measures ($\mu_{GC0.5}$and $\mu_{G0.5}$) are used inthis experiment. In the case of sounds classification, measuring the similarity rate by the $\boldsymbol{\mu_{Gc0.5}}$ permits to have an idea on the type of considered sound.

If $\boldsymbol{\mu_{Gc\,0.5}} \in$ [Threshold_minj, Threshold _maxj] then we identify the considered sound as type « **j** ».where,

$\text{Threshold\_min}_j = \min(\mu Gc_{0.5}(\text{speech}, \text{sound}_f))$

$\text{Threshold\_max}_j = \max(\mu Gc_{0.5}(\text{speech}, \text{sound}_j))$

Where, the word « sound » represents a set of sounds with the same type. And « speech » represents a set of reference speakers. J denotes a certain sound belonging to a certain class "j".

D. *Sinusoidal Parameters based audio classification [8]*

Speech signal contains both periodic and non-periodic information due to the impulsive nature of events or"noise-like" processes occurring in unvoiced regions. As a result, we can write for a time window segment of the underlying observed audio signal as follows:

$$y(n) = \sum_{l=1}^{L} a_l \cos(2\pi f_l n + \phi_l) + \varepsilon(n)$$

where n = 1,…,N_s is the sample index, θ= (a_l, f_l, ϕ_l) denote the sinusoidal parameters including the amplitude, frequency, and phase of the l-th sinusoidal component, respectively, L is the number of sinusoidal components in the signal, and ε (n) is the observation noise modeled as a zero-mean, additive Gaussian noise sequence. In general, it is of interest to estimate the corresponding sinusoidal parameters including frequencies f_l, amplitudes a_l, phases$\emptyset_l$ and the number of sinusoids, L. Figure 1[8] shows the basic block diagram of the process which is FDMSM model. Whereas, Figure 2 shows the block diagram of the proposed audio classification system [8].

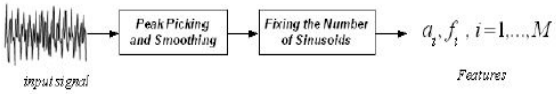


Figure 1

For extracting the sinusoidal parameters including amplitudes and their related frequencies in some Mel-bands, modified version of-the-state-of-the-art sinusoidal model also called Fixed Dimension Modified Sinusoidal Model (FDMSM) is employed [16].

At the end of the sinusoidal analysis using the FD-MSM approach we reach at triple features for each time window frame as:

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

$\varphi = (a_l, f_l, \phi_l)$

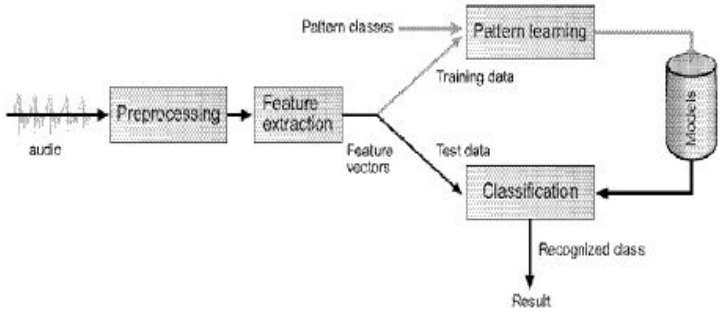


Figure 2

Further on Feature selection is done to find the smallest subset of the original features using an unsupervised method as pre-process.

IV. DISCUSSIONS ON RESULTS OBTAINED

A. *INDEPENDENT COMPONENT ANALYSIS [9]*

ICAexperimental results are shown in Figure 3[9].The observed mixture of signals is shown in Figure 4[9].The original speech signals are presented in Figure 3 and the mixed signals is shown in Figure4. The ICA algorithm is used to recover the data in Figure3 using only the data in Figure4.

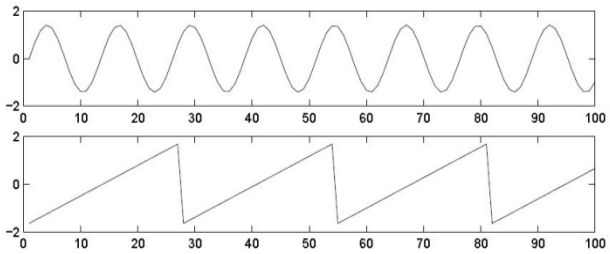


Figure 3

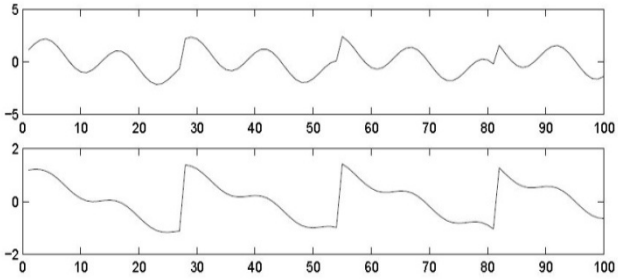


Figure 4

B. *HARMONIC STRUCTURE STABILITY ANALYSIS [4,5]*

In Figure 5[4,5], it can be seenthat the mixtures are well separated. The distance between music harmonic structures and the corresponding music AHS is small (the mean distances is 0.01 and 0.006 in experiment 1 and 2, respectively), and the distance between voice harmonic structures and the music AHS is bigger (the mean distances is 0.1 and 0.13 in experiment 1 and 2, respectively). So, the music AHS is a good model for music signal representation and for voice and music separation.

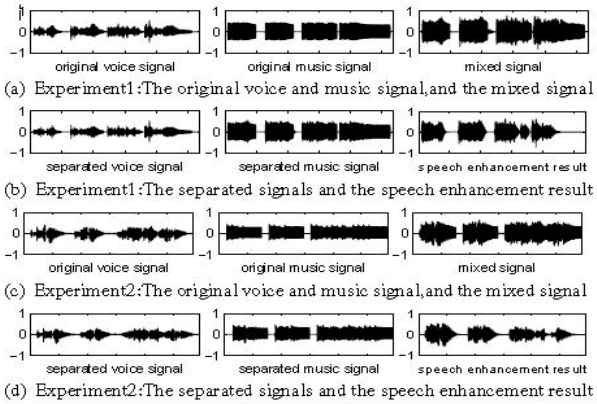
ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Fig. 5 also shows speech enhancement results obtained by speech enhancement software which tries to estimate the spectrum of noise in the pause of speech and enhance the speech by spectral subtraction.



(a) Experiment1:The original voice and music signal,and the mixed signal

(b) Experiment1:The separated signals and the speech enhancement result

(c) Experiment2:The original voice and music signal,and the mixed signal

(d) Experiment2:The separated signals and the speech enhancement result

Figure 5

C. *Second Order Statistical Measures Approach[7]*

This method has two aims: firstly, the discrimination between speech and other sounds; secondly the classification of the different acoustical sounds according to the μ_{Gc} values, used as a measure of similarity, corresponding for each sound. For example, if μ_{Gc} is within [2.52 – 4.92] then we can state that given sound should be music (Table 1).Table 1[7] shows Classification of sounds according to the μ_{Gc} range.

Table 1

μ_{Gc} range	**Probable classification**
0.13 – 0.27	Speech of a known speaker
0.27 – 0.78	Speech of a different speaker
0.78 – 1.8	Speech of a different speaker or human noise
1.8 – 2.46	Office noise
2.52 – 4.92	Music
8.23 – 10.34	Background noise

D. *Sinusoidal Parameters based audio classification[8]*

In this approach, assume that S(k) be the spectrum of current speech frame. To accomplish the peak picking process 8 msec hop size is chosen. It has fixed number of sinusoidal parameters and also preserve the synthesize quality as close as possible to the original signal in terms of perception. This results in the significant reduction as well as better clustering performance. Table 2[8] shows mixed type audio classification accuracy.

Table 2

Classifier	Accuracy	Parameters	Kernel
MLP	91.3	2 input 6 hidden layer neurons	-
SVM	93.79 97. 93 99.31 98.62	Linear γ=0.5 γ =0.3	- RBF RBF RBF
KNN	82.00 81.09	Max number of neighbors= 3 Max number of neighbors= 2	- -
RVM	97.14	γ = 3	RBF

As it was enlightened in simulation results, employing such sinusoidal parameters along with the so-called SVM

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

classifier, will successfully improve the performance of the proposed classification. Figure 6[8] is showing the decision boundaries using SVM[8].

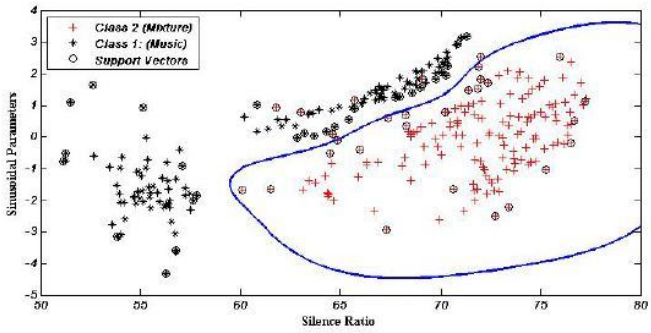


Figure 6

V. CONCLUSION

Signal separation is a difficult problem and no reliable methods are available for the general case. The detailed research on all the works done on separation of audio signals has led us to the conclusion that Harmonic Structure characteristic is the effective with respect to others as it preserves the audio quality. Moreover, most of the instrument sounds are harmonic in nature. Harmonic structure of music signal is stable. Harmonic structures of the music signals performed by different instruments are different. So, we can easily classify and separate different signals. However, this algorithm doesn't work for non-harmonic instruments, such as some drums. Some rhythm tracking algorithms can be used instead to separate drum sounds.

Pre-processing techniques like sinusoidal parameter approach and SOSM approach have also led us to the conclusion that selection of feature is an important step in the research field especially in music. That is why, the need to select the most efficient and accurate feature is essential for audio classification. Classifiers like SVM are proven to be the best classifiers as they more accurate than other although it may get trade off by the elapsed time.

REFERENCES

1. J. Pinquier, J. Rouas, and R. A. Obrecht, "Robust speech / music classification in audio documents," International ConferenceOn Spoken Language Processing (ICSLP), pp. 2005–2008, 2002.
2. G. R. Naik and D. K. Kumar "An Overview of Independent Component Analysis and Its Applications", Informatica 35,pp. 63–81,2011.
3. P. Vanroose, "Blind source separation of speech and background music for improved speech recognition", The 24th Symposium on Information Theory, pp. 103–108, May 2003.
4. Y. G. Zhang and C. S. Zhang "Separation of voice and music by Harmonic Structure Stabilty Analysis", Multimedia and Expo, ICME 2005, pp.562 - 565, 2005.
5. Y. G. Zhang and C. S. Zhang "Separation of Music Signals by Harmonic Structure Modeling",NIPS, 2005.
6. P. M. B. Mahalel, M. Rashidi, K. Faez, A. Sayadiyan" A New SVM-based Mix Audio Classification", NIPS, 2004.
7. H. Sayoud, S. Ouamour "Statistical Discrimination and Identification of Some Acoustic Sounds", GCC Conference (GCC),pp.1 – 5, 2006.
8. P. MowlaeeBegzadeMahale, A. Sayadiyan, and K. Faez "Mixed Type Audio Classification using Sinusoidal Parameters", 3rd International Conference on Information and Communication Technologies (ICTTA),2008.
9. A. Hyvärinen and E. Oja "Independent Component Analysis: Algorithms and Applications", Vol 13, pp, 411-430,2000.
10. A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval "Adaptation of Bayesian Models for Single-Channel Source Separation and its Applicationto Voice/Music Separation in Popular Songs",IEEE Transactions on Audio, Speech, and Language Processing, Vol.15 , Issue: 5,pp. 1564 – 1578,2007.
11. S. Koval, M. Stolbov, and M. Khitrov, "Broadband noise cancellation systems: new approach to working performance optimization," in EUROSPEECH' 99, pp. 2607–2610, 1999.
12. G. J. Jang and T. W. Lee, "A probabilistic approachto single channel blind signal separation," NIPS, 2003.
13. Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by independent component analysis," in ISMIR, pp. 281–282, 2002.
14. Anssiklapuri and Manuel Davy "Signal Processing methods for transcription", springer science, pp. 6-9, 2006.
15. F. BIMBOT, I. MAGRIN-CHAGNOLLEAU, et L. MATHAN 1995, "Second-Order Statistical measures for text-independent Broadcaster Identification". Speech Communication, Volume. 17, Number, 1-2, pp. 177-192, August 1995.
16. S. Z. Li, "Content-based audio classification and retrieval using the nearest feature line method", IEEE Transaction on Speech and Audio Processing, Vol. 8 , Issue: 5 , pp. 619 - 6252000.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

BIOGRAPHY

Silk Smitais a Masters in Engineering student in the Electronics and Communication Engineering Department, Birla Institute of Technology, Deemed University. Her research interests are music signal processing and signal processing.

SharmilaBiswasis PhD student in the Electronics and Communication Engineering Department, Birla Institute of Technology, Deemed University. Her research interests are music signal processing and signal processing.

Sandeep Singh Solankiis Associate Professor in the Electronics and Communication Engineering Department, Birla Institute of Technology, Deemed University. His research interests are music and speech signal processing and automation.