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. 3, Issue 3, March 2015

Study of Algorithms for Separation of Singing Voice from Music

Madhuri A. Patil[1], Harshada P. Burute[2], Kirtimalini B. Chaudhari[3], Dr. Pradeep B. Mane[4]

Department of Electronics, AISSMS's, College of Engineering, Pune, India[1]

Department of Electronics, AISSMS's, Institute of Information Technology, Pune, India[2]

Department of Electronics, AISSMS's, College of Engineering, Pune India[3]

Department of Electronics, AISSMS's, Institute of Information Technology, Pune, India[4]

ABSTRACT: Audio signal is an acoustic signal which has frequency range roughly in 20 to 20,000 Hz. Human auditory system has a wonderful ability of effectively focusing on sound in the surrounding. Most audio signals are from the mixing of several sound sources. Separation of singing voice from music has wide range of application such as lyrics recognition, alignment, singer identification, and music information retrieval. Music accompaniment that is often non-stationary & harmonic. Basically, audio signal is time frequency segments of singing voice. An audio signal classification system should be able to categorize different audio format like speech, background noise, and musical genres, singer identification, karaoke etc. In this paper, discuss about separation technique and classifier which are used for singing voice separation from music. Non-negative matrix factorization (NMF) is used for separation from music, Gaussian mixture model (GMM) & Support vector machine (SVM) classifier for the classification.

KEYWORDS: Singing voice separation, Classifier, Non-negative matrix factorization (NMF), Gaussian mixture model (GMM), Support vector machine (SVM)

I. INTRODUCTION

Audio signal separation is one of the complex areas of audio signal processing. In our surrounding there is number of audio signals with noise. Audio mixture is categories as music or speech & live recording or synthetic mixture. Separation is important for simultaneously separating multiple audio signals from mixture. Audio source signal are non-stationary. It is important to separate wanted signal & while reducing undesired interfering signals and noise, because human focus on one source effectively. Audio separation mainly depends upon techniques such as blind and inform source separation model. Multichannel processing has more source separation capability as compare to single channel. It is well known that the auditory system has a remarkable capability in separating sound from different sources. One important aspect of this capability is hearing out singing voice accompanied by musical instruments [13].There are wide range of application like melody extraction, singer identification, lyrics alignment and recognition, and content-based music retrieval etc [2]. Singer identification is another area for applying such type of system. In singing voice separation accuracy of singer identification is improved.

This paper explain, three popularly used techniques such as Non-negative matrix factorization (NMF), classification using Gaussian mixture model (GMM), and Support vector machine (SVM) which are used for separation of audio signal. NMF is unsupervised learning technique which is also called rank reduction technique used for single and multichannel source separation. GMM and SVM are used as classifier for the audio and speech signal separation.

The remainder of this paper is organized as follows. Section II presents literature survey to audio separation algorithms. Section III gives an overview of the algorithms. Section IV gives conclusion of the paper.

II. LITERATURE SURVEY

In 2013, Hiroshi Sawada and Hirokazu Kameoka, proposed new formulations and algorithms single channel NMF and multichannel NMF with considering its own spatial properties. Hiroshi derived multichannel algorithms where 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. 3, Issue 3, March 2015

single channel algorithms in number of iteration of convergence. Multichannel NMF with IS divergence better source separation than Euclidian distance for stereo music mixture [1].

In 2013, Bilei Zhu, Wei Li, Ruijiang Li, and Xiangyang Xue, proposed monaural singing voice separation by using new Non-negative matrix factorization (NMF) algorithm. NMF algorithm used to decompose long-window and short-window mixture spectrogram. NMF provide high quality separation for the singing voice and music. Algorithm performed very well when voice to music ratio is low or medium. By combining long and short algorithm gives good performance at high voice-to music ratio [2].

In 2010, Alexey Ozerov and Cedric Fevotte, gives brilliant idea about two interface methods as an Expectation-maximization (EM) algorithm for maximization of channels joint log-likehood and Multiplicative update (MU) algorithm for maximization of sum individual channel log-likehood complexity of grows linearly with number of model component [3].

In 2008, Olivier Gillet and Gael Richard, proposed novel system which is relies on combined used of classification and source separation. Proposed classifier as C-Support vector machine (C-SVM) whose generalization properties and discriminative power have been proved [5].

In 2013, Kun Han, and DeLiang Wang, proposed separation system by using Support vector machine (SVM) classification which required minimal training system trains SVM to provide initial classification and then used rethresholding to estimate ideal binary mask (IBM) [6].

In 2013, Ziqiang Shi, Jiqing Han, Tieran Zheng, and Ji Li, proposed two new generalization of mixture model such as, more general distance measure between two vector based on non-linear map which give more general mixture model called pseudo-GMM and another multiple different kind of distribution to interpret data from different sources. Ensemble classifier performed better as compare to using single component classifier [10].

In 2012, Elizabeth Godoy, Olivier Rosec, Thierry Chonavel, explain concept of voice conversion (VC) used GMM based transformation. Elizabeth proposed alternative to GMM as Dynamic frequency warping (DFW) for VC [11].

In 2007, Yipeng Li, and DeLiang Wang, used Gaussian mixture model (GMM) classifier for the detection of singing voice by using the feature of the database mixture [13].

III. ALGORITHMS OVERVIEW

Separation of signal techniques are used in many application NMF algorithms has wide range of applications such as image representation, document clustering, and music transcription [1]. Techniques that call factorization make use of the natural redundancy of the signal, mimicking human cognition which utilizes this redundancy to understand visual and audio signals [4]. NMF has been used to estimate clean speech from noisy observation. SVM is a state-of-the-art learning machine which is widely used for classification problems. Basically, SVM maximizes the margin of separation between different classes of training data, and as a result it shows good generalization. Utilize SVMs to produce initial classification boundaries and then derive new thresholds to classify T-F units in unseen acoustic environments [2]. Gaussian mixture model (GMM) is widely used in many application like pattern recognition, machine learning, and data mining and statistical analysis [10].

A. Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF) is more than 30 year old. Non-negative includes sparsity. Short-time Fourier Transform (STFT) to obtained complex value representation in the frequency domain. NMF imposes non-negative constrain which lead only additive, no subtractive combination of original data. NMF can use long-window and short-window spectrogram factorization, it can lead better performance of removing music interferences from singing voice [2]. NMF work as a part-based decomposition. NMF used to decompose the mixture spectrogram into set of component to different sound sources.

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. 3, Issue 3, March 2015

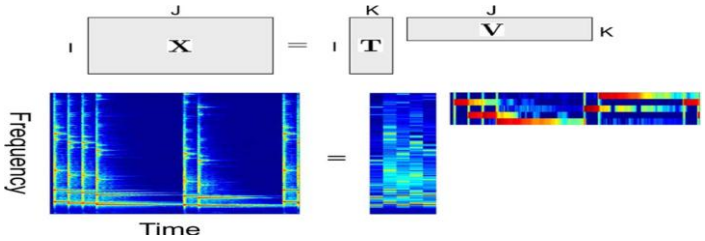

Fig.1 formulation of NMF (top) and its application to music signal (bottom). Frequent sound patterns are identified in a matrix **T** along with their activation period and strengths shown in matrix **V**. (reproduced from [1])

$$Xij = Tki\ Vkj \qquad (1)\ [1]$$

As the top part of fig.1 shows, NMF decomposes a given non-negative matrix **X** into two smaller non-negative matrices **T** and **V**. Observation matrix **x** is typically a phase-invariant time-frequency (T-F) representation (e.g. magnitude spectrogram or power spectrogram) of the input sound mixture, where is **I** the number of frequency channel and **J** is number of tine frames. Analysis of an audio/music signal with NMF, employ a short time Fourier transformer (STFT) to obtain complex/valued representations in the time-frequency domain [1]. Find distance/divergence of the matrix which is used in the NMF cost function. There are several choices available for the distance/divergence measures such as Euclidian distance, the generalized Kullback-Leibler (KL) divergence, and the itakura-saito (IS) divergence.

B. Support Vector Machine (SVM)

The support Vector Machine (SVM) was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research community. SVM is supervised learning method used for classification. SVM simultaneously minimize the imperial classification error and maximize the geometric margin. So SVM called maximum margin classifiers. Data is classified by using the hyperplane. Sample along the hyperplane called Support Vector (SV). The separating hyperplane is the hyerplane that maximize distance between the two parallel hyperplane [9].

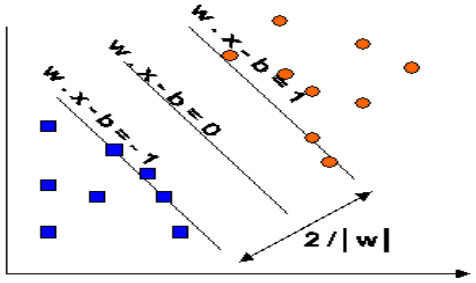

Fig.2 Maximum margin hyperplanes for a SVM trained with samples from two classes (reproduced from [9])

If distance or margin between parallel hyperplane is better than SVM gives good classification. Parallel hyperplanes can be described by equation

$$W \cdot x + b = 1 \qquad (2)\ [9]$$
$$W \cdot x + b = -1 \qquad (3)\ [9]$$
$$W \cdot x + b = 0 \qquad (4)\ [9]$$

In the fig.2 a separating hyperplane with the largest margin defined by $M = 2/|w|$ that is specifies support vectors means training data points closets to it. SVM used kernel function for classification of data [9]. When SVM output in one channel, then estimate parameters of a half-cauchy distribution to fit the output by maximal like-hood estimation [6].

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. 3, Issue 3, March 2015

C. Gaussian Mixture Model (GMM).

Separation of singing voice from music used Gaussian mixture model (GMM) as a classifier for the classification of the voice and unvoiced signal. Gaussian mixture model (GMM) is a mixture of several Gaussian distribution and can therefore represent different subclasses inside one class. GMM to represent perfectly the data distribution: the most important for classification is to obtain a good separator between the classes. This was confirmed by considering discriminative training of GMMs for classification. Gaussian mixture model (GMM) is supervised learning which is best on the maximum likelihood (ML) estimation using expectation maximization (EM). Compared traditional GMM with pseudo GMM the nonlinear maps have better performance on nonlinear problems, while the computational complexity is almost the same as the Expectation-Maximization (EM) algorithm for traditional GMM according to the iteration procedures [10]. In the training phase, a music database with manual vocal/nonvocal transcriptions is used to form two separate GMM: a vocal GMM and nonvocal GMM [12].

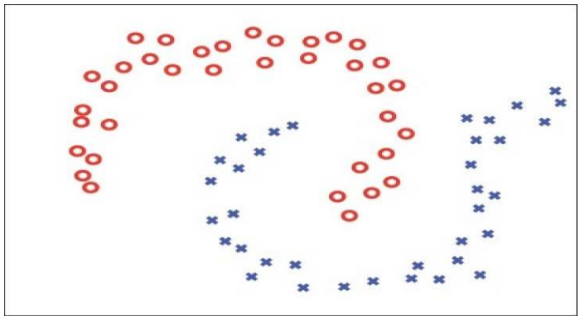

Fig.3 Data distribution that cannot well modeled by traditional GMM (reproduced from [10])

The expectation maximization (EM) algorithm is an iterative method for calculating maximum likelihood distribution parameter estimates from incomplete data. EM algorithm is high for two major reasons as similar to other kernel based methods, it has to calculate kernel function for each sample-pair over training set and in order to obtain the largest eigenvalue.

IV. CONCLUSION

In this paper, discussed about Non-negative matrix factorization (NMF) technique, classifier as Gaussian mixture model (GMM) and Support vector machine (SVM). Zhu [2] described the NMF technique gives high quality separation for both singing voice and music. Han and Wang [6] described that, for the speech separation SVM classifier gives more accurate classification as compare to GMM classifier.

REFERENCES

1. Hiroshi Sawada and Hirokazu Kameoka, Shoko Araki, and Naonori Ueda, "Multichannel Extensions of Non-Negative Matrix Factorization With Complex –Valued Data", IEEE Transactions on Audio, Speech, and Language processing, Vol. 21, no. 5, May 2013 .
2. Bilei Zhu, Wei Li, Ruijiang Li, and Xiangyang Xue, "Multi-Stage Non-Negative Matrix Factorization for Monaural singing Voice Separation", IEEE Transaction on Audio, Speech, and Language processing, Vol. 21, no. 10, October 2013.
3. Alexey Ozerov, and Cedric Fevotte, "Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation", IEEE Transaction on Audio, Speech, and Language processing, Vol.18, no.3, March 2010.
4. Romain Hennequin, Roland Badeau, and Bertrand David, "NMF With Time-Frequency Activation to Model Nonstationary Audio Events", IEEE Trasaction on Audio, Speech, and Language processing, Vol.19,no.4, May 2011.
5. Olivier Gillet, and Gael Richard, "Transcription and Separation of Drum Signals from Polyphonic Music", IEEE Transaction on Audio, Speech, and Language processing, Vol.16, no.3, March 2008.
6. Kun Hun, and DeLiang Wang , "Towards Generalizing Classification Based Speech Separation" , IEEE Transaction on Audio, Speech, and Language processing, Vol.21, no.1, January 2013.
7. Shi-Xiong Zhang, Mark J. F. Gales, "Structure SVMs For Automatic Speech Recognition" , IEEE Transaction on Audio, Speech, Language Processing, Vol.21, no.3, March 2013.
8. William M. Campbell, Joseph P. Campbell, Terry P. Gleason, Douglas L. Reynolds, and Wade Shen, "Speaker Verification Using Support Vector Machine and High-Level Features" , IEEE Transaction on Audio, Speech, and Language processing, Vol. 15, no. 7, September 2007.

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. 3, Issue 3, March 2015

9. Durgesh K. Srivastava, Lekha Bhambhu, "Data Classification Using Support Vector Machine", Journal of Theoretical and Applied Information Technology.
10. Ziqiang Shi, Jiqing Han, Tieran Zheng, and Ji Li, "Identification of Objectionable Audio Segments Based on Pseudo and Heterogeneous Mixture Model", IEEE Transaction on Audio, Speech, and Language processing, Vol.21, no.3, March 2013.
11. Elizabeth Godoy, Olivier Rosec, and Thierry Chonavel, "Voice Conversion Using Dynamic Frequency Warping with Amplitude Scaling, for parallel or Nonparallel Corpora", IEEE Transaction on Audio, Speech, and Language processing, Vol. 20, no.4, May 2012.
12. Wei-Ho Tsai, and Hao-Ping Lin, "Background Music Removal Based on Cepstrum Transformation For Popular Singer Identification", IEEE Transaction on Audio, Speech, and Language processing, Vol.19, no.5, JULY 2011.
13. Yipeng Li and DeLiang Wang, "Separation of Singing Voice From Music Accompaniment For Monaural Recordings", IEEE Transaction on Audio , Speech, and Language processing, Vol.15, no.4, May 2006.

.