International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Survey on "Singing Voice Separation and Singing Voice Classifier Technique for Voice Separation from Music Accompaniments"

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

[1, 2]Department of E&TC, Annasaheb Dange COE, Ashta, Maharashtra, India

Abstract: ***Singing voice separation from music is a kind of speech separation and it is a big challenge in many applications. Humanauditory system has a good ability of effectively focusing on sound in the surrounding. Most audio signals arefrom the mixing of several sound sources. Separation of singing voice from music has wide range of application suchas lyrics recognition, alignment, singer identification, and music information retrieval. Music accompaniment that isoften non-stationary & harmonic. Basically, audio signal is time frequency segments of singing voice. An audio signalclassification system should be able to categorize different audio format like speech, background noise, and musicalgenres, singer identification, karaoke etc. In this paper, discuss about different separation technique and classifier which are used for singing voice separation from music.***

Keywords: singing voice separation, singing voice classifier:-HMM, GMM, SVM

1. 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. The aim of speech separation is to separate the target speech from broadband or narrowband, periodic or a periodic background. Whereas, for singing voice separation, separating singing voice from broadband, periodic and correlated music 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].Singing voice separation is, in a sense, a special case of speech separation and has many similar applications. For example, automatic speech recognition corresponds to automatic lyrics recognition, automatic speaker identification to automatic singer identification, and automatic subtitle alignment which aligns speech and subtitle to automatic lyric alignment which can be used in a karaoke system. These applications also encounter similar problems. They perform substantially worse in the presence of background noise or music accompaniment.

2. Related Work

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 [8]. 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].

The several singing voice separation methods are as follows;
a) spectrogram factorization
b) model-based methods
c) Pitch-based methods.

Spectrogram factorization methods utilize the redundancy of the singing voice and music accompaniment by decomposing the input signal into a pool of repetitive components. Each component is then assigned to a sound source. Model-based methods learn a set of spectra from music accompaniment only segments. Spectra of the vocal signal are then learned from thesound mixture by fixing accompaniment spectra. Pitch-based methods use extracted vocal pitch contours as the cue to separate the harmonic structure of the singing voice.

These methods have their limitations. Spectrogram factorization methods encounter difficulties in assigning repetitive components or bases to the corresponding sound sources. The performance drops significantly when the number of musical instruments increases. Furthermore, it is difficult to separate singing voice from a short mixture since vocal signals typically have more diverse spectra than that of each instrument. Model-based methods require a considerable amount of music accompaniment only segments so that they can model the characteristics of the background music. Compared to spectrogram factorization and model-based methods, pitch-based methods potentially have fewer limitations. The only required cue is the pitch contours of the

singing voice which can be extracted from a very short mixture and does not need the accompaniment only parts.

3. Survey of Various Singing Voice Extraction Techniques:

A) Harmonic-locked loop technique:
In this system, the fundamental frequency of the singing voice needs to be known a priori. The system also does not distinguish singing voice from other musical sounds. When the singing voice is absent the system incorrectly tracks partials that belong to some other harmonic source. The harmonic-locked loop requires the estimation of a partials instantaneous frequency, which is not reliable in the presence of other partials and other sound sources. Therefore, the system only works in conditions where the energy ratio of singing voice to accompaniments is high [7].

B) Separation of singing and piano sound:
This system requires significant amount of prior knowledge, such as the partial tracks of premixing singing voice and piano or the music score for piano sound. This prior knowledge in most cases is not available. Therefore the system cannot be applied for most real recording [6].

C) Monaural speech segregation technique based on pitch tracking and amplitude modulation:
This system relies heavily on pitch to group segments. Therefore the accuracy of pitch detection is critical. This system obtains its initial pitch estimation from the time lag corresponding to the maximum of a summary of autocorrelation function. This estimation of pitch is unreliable for singing voice. This system assumes that voice speech is always present. For singing voice separation, this assumption is not valid. This system cannot separate unvoiced speech [5]

D)Computational auditory scene analysis (CASA):
Represents the first CASA attempt to musical sound separation. His system extracts onset and common frequency variation and uses them to group frequency partials from the same musical instrument together. However these two cues seem not strong enough to separate different sounds apart. The author suggested that other cues, such as pitch, should be incorporated for the purpose of sound separation. The pitch cue, or the harmonicity principle, is widely used in CASA systems.CASA is proposed in this method a lot of effort has been made to segregate speech from music accompaniments. But the performance of current CASA system is still limited by pitch estimation errors and residual noise [4].

E) Tandem algorithm:
Tandem algorithm used in voice speech segregation performs pitch estimation and voice separation jointly and iteratively. It is observed that the target pitch can be estimated from a few harmonics of the target signal. On the other hand, it can separate some target signals without perfect pitch estimation. This system show consistent performance improvement for all types of intrusion except rock music, presumably because of the strong harmonicity of the music Accompaniment. This indicates that separating speech from music is challenging to their tandem algorithm [3].

F) Tandem algorithm with trend estimation:
In the proposed work tandem algorithm is used to reduce pitch detection problem by using a trend estimation algorithm to bound the singing pitch contours in a series of time-frequency (T-F) blocks that have much narrower pitch ranges as compared to the entire possible range. The estimated trend substantially reduces the difficulty of singing pitch detection by eliminating a large number of wrong pitch candidates [1].

Detecting pitch values for singing voice in the presence of musicaccompaniment is challenging but useful for many applications. The pitch ranges are usually large to cover most ofthe possible pitches of singing voice such as from 80 Hz to 800 Hz. However, it is unlikely that pitch changes in such a wide range in ashort period of time. Furthermore, the upper pitch boundary of singing can be as high as 1400 Hz for soprano singers. It isdifficult to give an appropriate pitch range if no prior knowledge is given for an input song.

To address the above problems, we propose trend estimation Algorithm:

The advantages of adopting trend estimation are:
1) Appropriate pitch ranges of singing voice are estimated from the input instead of using a fixed range.
2) The trend is estimated dynamically so that it is able to fit the pitch trajectory of the corresponding singing voice.
3) It can be easily applied to improving singing pitch detection of existing systems by using a tighter pitch range for a given time frame.

Tandem algorithm which performs pitch estimation and voice separation jointly and iteratively. It is observed that thetarget pitch can be estimated from a few harmonics of the target signal. On the other hand, one can separate some target signals without perfect pitch estimation. Thus, their strategy is to have a rough estimate of the target pitch first and then separate the target speech by considering harmonicity and temporal continuity. The separated speech and the estimated target pitch are then used to improve upon each other iteratively. They show a consistent performance improvement for all types of intrusion except rock music presumably because of the strong harmonicity of the music accompaniment. This indicates that separating speech from music is challenging to their tandem algorithm.

Tandem algorithm detects multiple pitch contours and separates the singer by estimatingthe 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.

4. Survey of Various Singing Voice Classifier

a) Hidden Markov Model:
A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM can be presented as the simplest dynamic Bayesian network. In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden

Markov model, the state is not directly visible, but output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Note that the adjective 'hidden' refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly.

Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics. A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (or latent variables), which control the mixture component to be selected for each observation, are related through a Markov process rather than independent of each other. Recently, hidden Markov models have been generalized to pair wise Markov models and triplet Markov models which allow consideration of more complex data structures and the modelling of nonstationary data.

In the hidden Markov models, the state space of the hidden variables is discrete, while the observations themselves can either be discrete (typically generated from a categorical distribution) or continuous (typically from a Gaussian distribution). Hidden Markov models can also be generalized to allow continuous state spaces. Examples of such models are those where the Markov process over hidden variables is a linear dynamical system, with a linear relationship among related variables and where all hidden and observed variables follow a Gaussian distribution. In simple cases, such as the linear dynamical system just mentioned, exact inference is tractable (in this case, using the Kalman filter); however, in general, exact inference in HMMs with continuous latent variables is infeasible, and approximate methods must be used, such as the extended Kalman filter or the particle filter.

In singing voice separation, HMM is used to decode an input mixture into vocal and nonvocal sections. The signals after applying HPSS which attenuates the energy from music accompaniment is used instead of the original mixture. HMM can also be used for pitch extraction in musical accompaniment.

b) Gaussian mixture model

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. In the training phase, a music database with manual vocal/non vocal transcriptions is used to form two separate GMM: a vocal GMM and nonvocal GMM.

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 eigen value [9].

c) Support vector machine

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

There are two main reasons for using the SVM in audio classification. First, many audio classification problems involve high dimensional, noisy data. The SVM is known to behave well with these data compared to other statistical or machine learning methods. Second, the feature distribution of audio data is so complicated that different classes may have overlapping or interwoven areas. However, a kernel based SVM is well suited to handle such as linearly non-separable different audio classes. The classifier with the largest margin will give lower expected risk, i.e. better generalization. SVM transforms the input space to a higher dimension feature space through a nonlinear mapping function. Construct the separating hyper plane with maximum distance from the closest points of the training set. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

5. Conclusion

Singing voice separation from music accompaniments is a branch of speech separation process, which ongoing interesting research topic for many years, but still there is a lack in separating the signal from the mixture of signal with 100%accuracy. Many researches have been use many singing voice separation and singing voice classifier technique to separate singing voice from music accompaniments. A more efficient method is required for singing voice separation from music accompaniments.

References

[1] Chao-Ling Hsu, DeLiang Wang, Jyh-Shing Roger Jang, and Ke Hu, "A Tandem AlgorithmFor 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] Morales-Cordovilla, J.A., Peinado, A.M. ; Sanchez, V. ; Gonz.l ez, J.A., "Feature Extraction Based on Pitch Synchronous Averaging for Robust Speech Recognition", IEEE Transactions On Audio, Speech, and Language Processing, Vol. 19, No. 3, pp. 640 – 651, March 2011.
[3] Guoning Hu and Deliang Wang "A Tandem algorithm for pitch estimation and voice speechSegregation" IEEE Transaction on Audio, Speech, and Language Processing, VOL.18, NO. NOVEMBER 2010.
[4] Yipeng Li, DeLiang Wang, Separation of Singing Voice from Music Accompaniment for Monaural Recordings, IEEE Transactions on Audio, Speech, and Language Processing,v.15 n.4, p.1475-1487, May 2007.
[5] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitudeModulation," IEEE Trans. Neural Netw., vol. 15, no. 5, pp. 1135–1150, Sep. 2004.
[6] Y. Meron and K. Hirose, "Separation of singing and piano sounds," in Proc. 5th Int. Conf.Spoken Lang. Process. (ICSLP 98), 1998.
[7] A.L.C Wang. "Instantaneous and frequency-warped signal processing Technique for auditorySource separation."Ph.D. dissertation, Dept. Elect.Eng., Stanford Univ., Stanford, CA 1994.
[8] Chong Un; Shih-Chien Yang, "A pitch extraction algorithm based on LPC inverse filtering And AMDF", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 25, No. 6, pp. 565 – 572, 1977.
[9] 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.

Author Profile

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.