A SURVEY ON SINGING VOICE SEPARATION TECHNIQUES

Shamili P
PG Scholar: dept. of CSE
Vimal Jyothi Engineering College
Kannur, India

Ancy K Sunny
Asst. Professor: dept. of CSE
Vimal Jyothi Engineering College
Kannur, India

Dr. T. M Thasleema
Asst. Professor: dept. of CSE
Central University of Kerala
Kasargod, India

***Abstract*—** Extraction of singing voice from music signal is required for many applications such as MIR (Music Information Retrieval), automatic karaoke generator, interactive music player etc.. In many MIR studies like automatic lyrics recognition, identification of language of a song, automatic singer identification etc., has used the information on singing voice. But singing voice separation is a challenging problem. Since both singing voice and musical sounds are harmonic, simple harmonic extraction technique should not be used. And music signals do not satisfy the properties of noise. So classical noise suppression techniques will not work effectively. This paper is about various methods to separate singing voice from monaural music signals. Some technique uses models for singing voice separation; some others use properties of singing voice like high rankness, fluctuation etc.

Keywords—singing voice separation; Music Information Retrieval; harmonic; noise; low rankness; Fluctuation

I. INTRODUCTION

Singing voice enhancement is a challenging problem. There are many methods for singing voice enhancement. In most of the existing methods, an input signal is first transformed from time domain to time-frequency domain, and then singing voice is characterized there. Other components are suppressed with time-frequency masking. And finally, the estimated spectrogram of singing voice is transformed back to time- domain again. The important thing is how to distinguish singing voice from other components. This paper analyzes different techniques for singing voice separation for monaural music signals. Separating singing voices from music accompaniment is required for many applications such as music information retrieval (MIR) studies [1] like automatic lyrics recognition, identification of language of the song, automatic singer identification etc., interactive music player in which listener can adjust the volume of instruments used in the song and automatic karaoke generator. Singing voice separation is an interesting problem. Because there are some technical difficulties that make the singing voice separation is an interesting one. One of the difficulties is the similarity between singing voice and accompaniments, e.g., a piano, a guitar, and percussions. That is both the spectra of singing voice and harmonic instruments, such as a piano and a guitar, have harmonic structure. Therefore it is difficult for a simple harmonics extraction technique to detect only the singing voice in music signals. Second difficulty is that accompanying instruments do not satisfy some properties of "noise" such as whiteness and stationarity. Since music signals are not white noise and nor stationary, classical noise suppression technique will not work effectively in singing voice separation.

For singing voice separation, speech separation technique cannot be effective. Because the difference between singing voice and speech are significant. The main difference is the presence of an additional formant, known as the singing formant. This singing formant helps the voice of a singer to stand out from the accompaniment. However, the singing formant does not exist in other types of singing such as the ones in rock and country music. Another difference is based on the way of singing and speech is uttered. During singing, a singer often intentionally stretches the voiced sound and shrinks the unvoiced sound to match other musical instruments. This has two results. First, it changes the percentage of voiced and unvoiced sounds in singing. The majority of sounds generated during singing is voiced (about 90%), while speech has a larger amount of unvoiced sounds. Second, the pitch dynamics (the evolution of pitch in time) of singing voice tends to be piece-wise constant with abrupt pitch changes in between. This is opposite to the declination phenomenon in natural speech where pitch frequencies slowly drift down with smooth pitch change in an utterance. Besides these things, singing voice also has a wider pitch range. The pitch range of normal speech is between 80 and 400 Hz while the upper pitch range of singing can be as high as 1400 Hz. From the sound separation point of view, the major difference between singing and speech is the nature of other concurrent sounds. In a real acoustic environment, speech is usually altered by interference that can be harmonic or non-harmonic, narrowband or broadband. In most cases the interference is independent of speech in the sense that the spectral contents of target speech and interference are uncorrelated. For recorded singing voice, however, it is almost always accompanied by musical instruments that in most cases are harmonic, broadband, and are associated with singing since they are composed to be a coherent whole with the singing voice. These differences make the singing voice separation from the accompaniments potentially more challenging.

There are many methods for singing voice separation. Some technique uses models for singing voice separation; some others use properties of singing voice like low rankness, fluctuation etc. The rest of this paper is organized as follows.

ISSN: 2278 – 1323

The section II briefly describes various methods for singing voice separation. The section III gives a performance comparison of these methods. Finally, section IV concludes the paper.

II. SINGING VOICE SEPARATION TECHNIQUES

A song is a combination of singing voice and background music. Let x(n) be a mixture containing voice signal v(n) and a music signal m(n), where n is a discrete time index (x(n) = v(n) + m(n)). The problem is to estimate the voice contribution v(n) in the observed signal x(n). Following are some of the techniques for singing voice separation from a song.

A. GMM Based Source Separation

This is a model based technique proposed by A Ozerov et al. [2] that uses Gaussian Mixture Model (GMM) to separate singing voice. First the mixed signal x is transformed from time domain to time frequency domain using Short Time Fourier Transform (STFT). There are three steps in GMM based source separation.

- GMM sources modeling: There are two sources in a song, ie., the voice to be extracted and background music. In this step the short time Fourier spectra V_t of voice signal v and M_t of music signal m at time t are modeled with a GMM.
- Separation by adaptive Wiener filtering: Source separation is performed in the Short Time Fourier Transform (STFT) domain with adaptive Wiener filtering. This technique separates mixed signal into music and singing voice.
- Inverse Fourier Transform: It is performed to convert the signal back to time domain again.

B. Source Adapted Models

In this method the short time Fourier spectra of mixed signal are segmented into vocal and non-vocal frames [2]. Music model are learned from non-vocal frames and voice model are learned from vocal frames in a maximum likelihood manner. Then adaptive wiener filtering is performed for source separation. Joining of vocal and non-vocal frames is performed to estimate voice signal. Then inverse Fourier transform is performed to obtain signal back to time domain again.

Drawback of this method includes the following

- Low quality sound
- Required that the song should already be manually segmented into vocal and non-vocal frames by user
- Training sources similar to those to be separated are needed to achieve a satisfactory performance

C. Bayesian Models

The Bayesian model proposed by A. Ozerov et al. [3] is an adapted model which aims at adjusting a posteriori the models by tuning their characteristics to those of the sources actually observed in the mix can be used.

In popular songs, there are usually some portions of the signal when music appears alone. Corresponding temporal segments are known as non-vocal parts in contrast to vocal parts, i.e., parts that include voice. The key idea in this adaptation scheme is to use the non-vocal parts for music model adaptation. Then, the obtained music model and the general voice model are further adapted on the totality of the song. This method consists of the following three steps.

1) The song X is segmented into vocal parts and non-vocal parts.
2) An acoustically adapted music model is obtained from the non-vocal parts
3) The acoustically adapted music model and the general voice model are further adapted on the entire song X.

Advantages

- Adaptation scheme can consistently improve the separation performance
- Yields on average a 5-dB improvement which bridges half of the gap between the uses of general models on the one hand and ideal models on the other hand.
- Automatic separation of vocal or non-vocal segments

Disadvantages

- Quality of separated singing voice is low.

D. Pitch Based Inference

This is a systematic approach to identify and separate the unvoiced singing voice from the music accompaniment proposed by Hsu and Jang [4]**.** There are three stages. Stage 1 takes a mixture of singing voice and music accompaniment as input and is responsible for Accompaniment/Unvoiced singing voice/Voiced singing voice (A/U/V) detection and time–frequency decomposition. Stage 2 identifies voiced-dominant T-F units within each voiced frame and unvoiced-dominant T-F units within each unvoiced frame. In stage 3, the voiced- and unvoiced-dominant T-F units are re-synthesized independently and are then combined to form the separated singing voice.

A/U/V Detection: This block employs a continuous hidden Markov model (HMM) to decode the mixture input into accompaniments, unvoiced, and voiced segments

Time–Frequency Decomposition: The input song mixture is decomposed into 128 channels using the gammatone filterbank whose center frequencies are quasi-logarithmically spaced from 80 Hz to 5 kHz. These signals in different frequency channels are then split up into overlapping frames. These frames are sensory elements called time frequency (T-F) units because they are indexed by both time and frequency. After this the input signal is decomposed into small sensory elements and is ready to be grouped.

Unvoiced-dominant T-F unit identification within unvoiced frames: here classification of T-F units within unvoiced frames into either unvoiced-dominant or accompaniment dominant is performed. The energy of the T-F units of the singing voice is compared with that of the corresponding T-F units in music instruments. If the first one is larger the T-F unit is labeled as unvoiced-dominant, otherwise it is labeled as accompaniment dominant.

Advantages

- Quality of separated singing voice is improved
- Deal with the problem of separating unvoiced singing
- Complete framework for pitch based inference method to separate singing voice

Disadvantages

ISSN: 2278 – 1323

- Pitch based inference has their pros and cons. And require a method that takes the pros of the method and alleviate its cons.

E. Robust Principal Component Analysis

This technique, proposed by P.S. Huang et al. [5] is based on the idea that repetition is a core principle in music; therefore it can be assumed that the music instruments lie in a low-rank subspace. But the singing voice has more variation and is relatively sparse within a song. Based on these assumptions, Robust Principal Component Analysis (RPCA), which is a matrix factorization algorithm, is used for solving underlying low-rank and sparse matrices.

Cand`es et al. [6] proposed RPCA, which is a convex program, for recovering low-rank matrices when a fraction of their entries have been corrupted by errors, i.e., when the matrix is sufficiently sparse. Since music instruments can reproduce the same sounds each time they are played and music has, in general, an underlying repeating musical structure, music accompaniments can be assumed as a low-rank signal. Singing voices have more variation (higher rank) but are relatively sparse in the time and frequency domains. Singing voices are the components making up the sparse matrix. By RPCA, the low-rank matrix L contains music accompaniment and the sparse matrix S contain vocal signals. The separation can be done as follows: First, compute the spectrogram of music signals as matrix M, obtained from the Short-Time-Fourier Transform (STFT). Then use the inexact Augmented Lagrange Multiplier (ALM) method, which is an efficient algorithm for solving RPCA problem, to solve L + S = |M|, given the input magnitude of M. Then by RPCA, obtain two output matrices L and S. Given the separation results of the low-rank L and sparse S matrices, further apply binary time-frequency masking methods for better separation results. Once the time-frequency mask is computed, it is applied to the original STFT matrix M to obtain the separation matrix $X_{singing}$ and X_{music}.

Advantages

- Without using prior training or requiring particular features, this technique can achieve around 1-1.4 dB higher GNSDR compared with previous state-of-the-art approaches, by taking into account the rank of music accompaniment and the sparsity of singing voices.

F. Based on Repetition

Repetition is the basis of music as an art. A typical piece of popular music has generally an underlying repeating musical structure, with distinguishable patterns periodically repeating at different levels, with possible variations. An important part of music understanding is the identification of those patterns. On this basis, Z. Rafii and B. Pardo [7] proposed a simple method for separating music and voice, by retrieval of the repeating musical structure. First, the period of the repeating segment structure is found. Then, the spectrogram is decomposed at period boundaries and the segments are averaged to create a repeating segment model. At last, each time-frequency bin in a segment is compared to the model, and the mixture is segmented using binary time-frequency masking by labeling bins similar to the model as the repeating background.

To identify the repeating segments in a song, it is required to estimate a period of the repeating musical structure. Periodicities in a signal can be found by using the autocorrelation, which measures the similarity between a segment and a lagged version of itself over successive time intervals. After calculating the period p of the repeating musical segment, use it to evenly segment the spectrogram into segments of length p. Then compute a mean repeating segment over the segments of V, which can be thought of as the repeating segment model. The key idea is that time-frequency bins containing the repeating structures would have similar values at each period, and would also be similar to the repeating segment model. After computing the mean repeating segment, each time-frequency bin is divided in each segment of the spectrogram by the corresponding bin in repeating segment model spectrogram. After this take the absolute value of the logarithm of each bin to get a modified spectrogram where time-frequency bins repeating at period p have values near 0. V can then be partitioned by assigning time-frequency bins with values near 0 in modified spectrogram to the repeating background. This assumes that the repeating structure (the music) and the varying sound (the vocals) have sparse and disjoint time frequency representations. But practically time-frequency bins of music and voice can overlap, and furthermore the repeating musical structure generally involves variations.

Advantages

- It can achieve better separation performance than an existing automatic approach, without requiring particular features or complex frameworks.
- Simple, fast and completely automatable.

Disadvantages

- Better repeating period finder is required

G. Two-Stage Harmonic Percussive Sound Separation

Harmonic Percussive Sound Separation (HPSS) [8] is a singing voice enhancement technique. HPSS algorithm separates a signal into two components: harmonic and percussive. H. Tachibana et al. [9] used this algorithm for singing voice separation by exploiting two differently resolved spectrograms; one has rich temporal resolution and poor frequency resolution, while the other has rich frequency resolution and poor temporal resolution. This technique is based on the fluctuation of singing voice. There are three components in a song: harmonic (stationary, sustained), percussive (non-stationary, transient), singing voice (Fluctuated, quasi-stationary). Spectrogram with rich temporal resolution and poor frequency resolution separates percussive component from the song. The remaining components are singing voice and harmonic component. The spectrogram with rich frequency resolution and poor temporal resolution separates singing voice from harmonic components

Advantages

- Improved separation performance.

Disadvantages

- This technique will be ineffecient when singing voice was not fluctuating sufficiently and when accompanying sounds were fluctuating.

ISSN: 2278 – 1323

III. COMPARITIVE STUDY OF VARIOUS METHODS

The GMM-based approach provides the advantage of being sufficiently general and applicable to a wide variety of audio signals. These methods have shown good results for the separation of speech signals and some particular musical accompaniments. The basic idea behind these techniques is to represent each source by a GMM, which is composed by a set of characteristic spectral patterns. Each GMM is learned on a training set, which contains samples of the corresponding audio class (for instance, speech, music, drums, etc.). These models can be seen as general or a priori models, as they are supposed to cover the range of properties observable for sources belonging to the corresponding class. An efficient model must be able to yield a rather accurate description of a given source or class of sources, in terms of a collection of spectral shapes corresponding to the various behaviors that can be observed in the source realizations. This requires GMMs with a large number of Gaussian functions, which creates a number of problems. In addition to this the performance of general models is rather poor. It may be more efficient to use adapted models, i.e., models with characteristics close to those of the mixed sources.

Bayesian models overcome the above difficulties. This approach improves the quality of the source model, while keeping its dimensionality reasonable. The goal of model adaptation is to replace the general models (which match well the properties of the training sources, but not necessarily those of the corresponding sources in the mix), with adapted models adjusted so as to better represent the sources in the mix, thus leading to an improved separation ability.

The pitch-based inference methods use vocal pitch contours to extract the harmonic structures of singing voice. It therefore avoids the difficulties of clustering in the presence of multiple instruments and the length limitation of non-vocal segments in a song. Separating unvoiced sound is more difficult than separating voiced sound for two reasons. First, unvoiced sound lacks harmonic structure and is often acoustically noise-like. Second, the energy of unvoiced sound is usually weaker than that of voiced sound. As a result, unvoiced sound is more susceptible to interference from the accompaniment.

Robust Principal Component Analysis can achieve around 1∼1.4 dB higher GNSDR compared with previous technique without using prior technique. The separation technique that uses repetition of musical structure is the simplest one. Repetition is a important principle in music. This is true for popular songs, generally contain a noticeable repeating musical structure, over which the singer performs varying lyrics. Among these techniques 2-stage HPSS can provide better results. It uses temporal variability of singing voice. Although this technique require long time to separate singing voice. Because it is required to apply the technique twice using two differently resolved spectrogram. So a faster method for singing voice separation is required.

IV. CONCLUSION

Singing voice separation is a branch of speech separation process, which is an ongoing interesting research topic for many years, but still there is a lack in separating the required signal from the mixture of signals with 100% accuracy and be used by the common people. Many researchers have been done in various ways using the parameters like pitch, phase, magnitude, amplitude, frequency and energy, spectrogram of the speech signal. The pros and cons of each separating process are surveyed here. A more efficient method is required for singing voice separation from monaural music signals.

REFERENCES

[1] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music information retrieval from a singing voice using lyrics and melody information," EURASIP J. Adv. Signal Process., 2007, Article ID: 038727.

[2] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. WASPAA*, 2005, pp. 90–93.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[4] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1 K dataset," IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 2, pp. 310–319, Feb. 2010.

[5] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. H. Johnson, "Singing voice separation from monaural recordings using robust principal component analysis," in Proc. ICASSP, 2012, pp. 57–60.

[6] Emmanuel J. Cand`es, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," J. ACM, vol. 58, pp. 11:1–11:37, Jun. 2011.

[7] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in Proc. ICASSP, 2011, pp. 221–224.

[8] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, "Comparative evaluations of multiple harmonic/percussive sound separation techniques based on anisotropic smoothness of spectrogram," in *Proc. ICASSP*, 2012, pp. 465–468.

[9] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 1, january 2014

ISSN: 2278 – 1323