International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 11 | Nov 2020 www.irjet.net p-ISSN: 2395-0072

Review of Vocal Separation from Mixed Music Signals

Anila R[1], Safoora O K[2]

[1]*M. Tech Student, Department of Electronics & Communication Engineering, College of Engineering Thalassery, Kannur, Kerala, India*
[2]*Assistant Professor, Department of Electronics & Communication Engineering, College of Engineering Thalassery, Kannur, Kerala, India*

Abstract - *This study has been undertaken to identify different methods of vocal identification and its extraction techniques from a mixed music signal or simply from a master track. The identification and the extraction process of the vocal from the mixed music signal is complex because some musical instruments like Saxophone, have the sound which resembles with the voice of human.*

For the identification of the signal it requires some features of music signals like Mel Frequency Cepstral Coefficient (MFCC),Perceptual Linear Predictive Coefficient(PLP), Log Frequency Power Coefficient(LFPC),etc. Using these features the data signal or the vocal signal can be extracted from the input music signals.

***Key Words*: Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Predictive Coefficient(PLP), Log Frequency Power Coefficient(LFPC), K-nearest neighbour(K-NN), Support Vector Machine(SVM).**

1. INTRODUCTION

Vocal detection refers to the task that detect or identify the vocal sounds or the human singing voice from the input mixed music signals. Music is the rhythmic signal which can be the combination of one or more human voice and can also contain one or more musical instruments plays behind just to support the human singing. Our aim is to identify each signal source and to separate them. For this purpose first we want to identify the musical features such as MFCC,PLP,LFPC, pitch, tone, frequency, etc from the music signal. Most of the studies in this area are based on the spectrogram of the music signal. For the spectrogram calculation the STFT is used and then these spectrogram are used further for the classification, identification and the extraction criteria. By the extraction of the source we can automatically get the karaoke, and is also applicable in forensic area to identify the person.

The rest of the paper is organized as follows. Literature Review in section II. Application in section III. Concluding remarks are given in section IV.

2. LITERATURE REVIEW

In this system the wave file is given as the input music signal, this is given to the vocal and non-vocal classifier section and it will classify both vocal and non vocal signals separately, then it is fed to STFT. In STFT, the spectrogram of the musical signal is the output and is given to source separation and then it will identify only the vocal track or the vocal signals only and the output of the system will be only the desired vocal track present in the given input. It can be described easily using a block diagram:

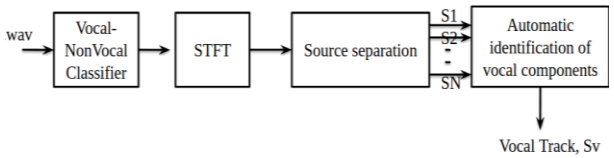

Figure 1:- Block diagram of vocal source separation system.

Bhiksha Ray and others proposed a system for separating a foreground singer from the background music. The main difference between this system and the system proposed by the Shanker Vembu is that, in this system the vocal and non-vocal areas in the music signal is not get separated automatically. Here we want to assume that the area to be identified is given or the user want to choose it manually, while the previous paper detect and identify the areas of vocal automatically. The main advantage of this system is that the user can modify the input music signal such as pitch, harmonization, etc. here distribution function, conditional probability function is calculated. For each component signal of the mixture, each frequency magnitude are learned from a separate unmixed training recording. Finally we get that probability of voice present in musical signal plus probability of background music present in the mixed musical signal. So we can separate it easily.

In [3], Derry Fitz Geralg introduce a novel vocal separator system is introduced which find k-nearest neighbour to each frame of a spectrogram of the input music signal. As common, here also magnitude spectrogram is finding out using STFT. Assume vocal signal is sparse and non-repeating. This paper follows algorithm model. After finding the magnitude spectrogram a distance matrix is calculated. The distance matrix is defined as the squared Euclidean distance between the *frame.*

$D_{k,l}=\sum(X_k-X_l)^2$ Where X is the mixture magnitude spectrogram which has the size of n*m. Xk is the kth spectrogram frame of the magnitude spectrogram and Dk,l is the squared Euclidean distance between from k and l.

Calculating the distance between all frame results in a symmetric matrix D of size n*m. this matrix is then sorted in to ascending order, hence we obtain another matrix P. the background music estimated for the kth frame is $Y_k=\mu(P)$

Where , Yk is the kth frame of the estimated background music spectrogram and μ denotes median operat Here we are assuming that the background music cannot have the greater energy at a given time-frequency bin than that of the mixture signal. Therefore, eliminate Y values which have the greater value than those of the original mixtur $Y_{f,k}=min(X_{f,k},Y_{f,k})$ A binary mask could be obtained for separating background music and voice signal by comparing the values in X with those in the Y. Here Gaussian radial basis function approach is used $W_{f,k}=\exp(-(\log X_{f,k}- \log Y_{f,k})^2/2\lambda^2)$ The complex values will be the result of this. Then the complex valued background music spectrum B can be estimated as $B=W \otimes R$ The background music signal can be recovered by taking the inverse STFT of B. Similarly, the complex valued spectrogram of voice signal can be obtained by, V=(1-W)⊗ R

The voice signal can be recovered by taking the inverse STFT of the signal V. In [4], the Zafar Rafii and Bryan Pardo also uses the algorithmic method. The main difference between[3] and [4] is that, in [3] the periodicity is not taken into account while in[4] the periodicity is also considered. Here also at first the spectrogram is calculated. Denoting the spectrogram of mixed music signal as V. Then the autocorrelation of each frequency component is obtained and form a correlation matrix, referred it as matrix B. By taking the mean over the rows of the autocorrelation matrix, obtain the vector, b which estimates the overall acoustic self-similarity of the mixture as a function of the time lag. Then the obtained vector is normalized by its first coefficient. This idea is similar to the beat-spectrogram. Beat-spectrogram is to characterize the rhythm of the audio signal.

The period P of the repeating musical structure is then defined as the period of the longest strong repeating pattern in mixture. After estimating P, then the spectrogram is segmented evenly into length P. then mean repeating segment ,

$$\overline{V}(i,l)=\left(\prod_{k=1}^{r} V(i,l+(k-1)\,p)\right)^{\frac{1}{r}}$$

for i=1,2,.....,n and l=1,2,....,p

Then dividing the each time frequency bin in each segment of the spectrogram then take the absolute value of the logarithm of each bin to get a modified spectrogram ,

$$\widetilde{V}(i,l+(k-1)\,p)=\left|\log\left(\frac{V(i,l+(k-1)\,p)}{\overline{V}(i,l)}\right)\right|$$
$$i=1\ldots n,\ l=1\ldots p \text{ and } k=1\ldots r$$

But in practical case, the time-frequency bins of music and vocal can overlap, therefore a binary time-frequency mask M is to be calculated.

$$M(i,j)=\begin{cases}1 & \text{,if } \widetilde{V}(i,j)\le t\\ 0 & \text{,otherwise}\end{cases}$$

Once M is calculated, it is symmetrized and applied to STFT of the mixture to get the STFT of the music Xmusic and STFT of voice Xvoice can be calculated,

$$\begin{cases}\hat{X}_{music}(i,j) &= M(i,j)\,X(i,j)\\ \hat{X}_{voice}(i,j) &= (1 \quad M(i,j))\,X(i,j)\end{cases}$$

This method has the advantage that it is being simple, fast and completely automatable and the disadvantage is the pitch, time and multidimensional information of the music signal is not taken into account.

In [5], it is the continuation of the previous paper[4], the authors continue the previous paper and introduced another way of separation. This method is a novel and simple approach for separating the background from the non-repeating foreground in a mixture.

The repeating segment models S(i,l),

$S(i,l)=median\{ V(i,l + (k-1)p) \}$

for i=1,2,...n(frequency) and l=1,2,...p(time), where p is the period length and r is the segment.

This is the element wise median of the r-segment. Once S is obtained, it is used to derive repeating spectrogram of model, W, by taking element wise minimum between S and each of the r-segment of the spectrogram V.

W(i,l+(k-1)p)=min { S(i,l), V(i,l+(k-1)p) }

for i=1,...n , l=1,2,...p and k=1,2,..r.

Then soft time frequency mask can be calculated by normalizing the repeating spectrogram model by the spectrogram V, element wise.

M(i,j)=W(i,j) / V(i,j)

with M(i,j) belongs to [0,1],

for i=1,2,...n(frequency) and j=1,2,...m

The time frequency mask M is then symmetrized and applied to the STFT ,x of the mixture x. the estimated music signal is obtained by inverting the resulting STFT into the time domain. The estimated voice signal is obtained by simply subtracting the time domain music signal from the mixture signal.
In [6], Bernhard Lehner and others proposed a system Online loudness-invariant vocal detection in mixed music signal. Here MFCC is utilized and also the version of MFCC is used, MFCC delta. SVM classifier is used for vocal and non-vocal separation. The output will be obtained according to the input music signal. In this paper the authors describes about the three treats.

1. False positive.
2. Low SNR
3. Dataset effect.

False positive means highly harmonic instruments have an increased risk for misclassified as vocal. ie, some musical instruments produces sounds which is similar to the human voice.

Low SNR refers that many recordings are produced with less than the optimal recording equipment and mixing, mastering skills etc, often with a low SNR. Here SNR is also described as Vocal to accompaniment ratio.

Dataset effect, music is a vast area and each region has its own particular type of music. So for the better result the dataset should be created in such a way that it is applicable for all music types. More the data's are collected and trained the system, more will be the result will be obtained.

In [7], Rupak Vignesh Swaminathan and Alexander has introduced a system which improves the voice separation using the attribute aware of deep learning network. The system functions are shown in the block diagram given below,

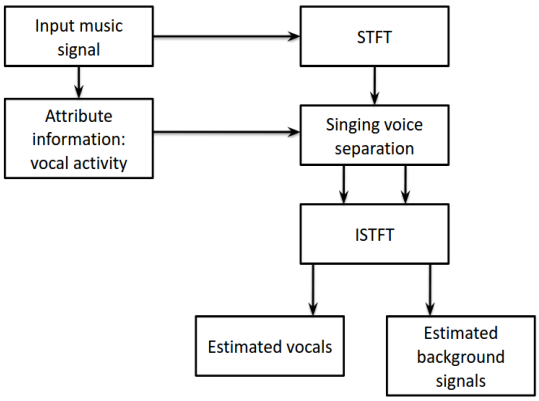

Figure 2: Block diagram of improving singing voice separation using attribute aware deep network.

The input mixed music signal is given as the input an a part of the input signal is given to STFT and also to Attribute information block. In the STFT block the signal is converted to the time frequency domain. In the attribute information Oracle label is used. The labels are represented as a one-hot encoder vector of two dimensions. The signing voice detections closely related to the timber classification or the instrumental recognition. Here CNNs for learning the singing voice characteristics and train it to output vocal activity predictions which fed in to the SVS network, ie, singing voice separation block. Then the inverse STFT is taken in order to obtain the voice and background music signal in time domain. The main problem about this system is that it is only applicable to some attributes. Prithish chandna and others introduced a system, content based singing voice extraction from a musical mixture. This system follows an encoder-decoder architecture and taken as input the magnitude component of the spectrogram of a musical mixture with vocals The encoder part of the model is trained via knowledge distillation using a teacher network to learn a content which is decoded to generate the corresponding vocoder features. The main advantage of this system is that the user can also extract the unprocessed raw vocal signal from the mixture even it was not the processed mixture dataset with singer not seen during the training time. Here the AutoVC model is used as the teacher network for knowledge distillation to train a network to learn the content from an input mixture spectrogram. A decoder is then trained to generate vocoder features given this content embedding which are then used for vocal synthesis. The original AutoVC model uses the mel-spectrogram as a representation of the vocal signal, but here they used vocoder feature. The reason to use the vocoder feature is that it is easier that the mel-spectrogram feature. The encoder and decoder of the network both take the singer identity as a one-hot vector.

3. APPLICATIONS

By extracting the vocals from the mixed music signals, we can easily separate or the user can create the karaoke of the song automatically, singer identification, automatic alignment of the lyrics to music. User can also remix and sample for use in new composition. It can also be used in the forensic area to identify the particular person's voice using the huge data stored in its database.

4. CONCLUSION

In this paper we have discussed the different methods for the vocal extraction from the input music signal. The spectrogram method are utilized in most of the studies while some papers are based on the non repeating structure of the music. The vocal extractions are done using the k-nn and also using SVM. This study is done to improve the further steps in the vocal extraction area.

REFERENCES

[1] Bernhard Lehner, Jan Schluter and Gerhard Widmer,"Online , Loudness-invariant Vocal Detection in Mixed Music Signal", IEEE/ACM Transactions on Audio, Speech and Language processing, VOL.XX, NO.X, XX 2018

[2] Derry Fitz Gerald, "Vocal Separation using Nearest neighborhood and Median Filtering"ISSC 2012,NUI Maynooth, June 2012

[3] Rupak Vignesh Swaminathan, Alexander Leech,"Improving Singing Voice Separation using attribute aware deep network", IEEE-2019, International workshop on Multilayer Music Representation and Processing (MMRP).

[4] Pritish Chandna, Merlijn Blaauw,Jordi Bonada,Emilia Gomez, "Cobtent Based Singing voice Extraction from a musical mixture", arXpv.2002.04933V2[eesa.AS],Feb-2020.

[5] S. Vembu and S. Baumann," Separation of Vocals from polyphonic audio recordings", in Proceedings to 6tg International Conference on Music Information Retrieval, London, UK , pp 337-344, September 11-15,2005.

[6] B. Raj, P. Smaragdis, M. Shashanka, and R. Singh, "Separating a foreground singer from Background music," in Proceedings to International Symposium Frontiers of Research on Speech and Music, Mysore, India, May 8–9, 2007.

[7] Zafar Rafii and Bryan Pardo," REpeating Pattern Extraction Technique (REPET): A Simple Method for Music/Voice Separation", IEEE Transaction on Audio, Speech and Language Processing, VOL. 21, NO. 1, JANUARY 2013.

[8] Zafar Rafii and Bryan Pardo," A Simple Music/Voice Separation Method Based on the Extraction of the Repeating Musical Structure", Conference Paper in Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on · June 2011