An Effective Framework for Speech and Music Segregation

Sidra Sajid, Ali Javed, and Aun Irtaza
Department of Software Engineering, University of Engineering and Technology Taxila, Pakistan

Abstract: *Speech and music segregation from a single channel is a challenging task due to background interference and intermingled signals of voice and music channels. It is of immense importance due to its utility in wide range of applications such as music information retrieval, singer identification, lyrics recognition and alignment. This paper presents an effective method for speech and music segregation. Considering the repeating nature of music, we first detect the local repeating structures in the signal using a locally defined window for each segment. After detecting the repeating structure, we extract them and perform separation using a soft time-frequency mask. We apply an ideal binary mask to enhance the speech and music intelligibility. We evaluated the proposed method on the mixtures set at -5 dB, 0 dB, 5 dB from Multimedia Information Retrieval-1000 clips (MIR-1K) dataset. Experimental results demonstrate that the proposed method for speech and music segregation outperforms the existing state-of-the-art methods in terms of Global-Normalized-Signal-to-Distortion Ratio (GNSDR) values.*

Keywords: *Ideal binary mask, source segregation, repeating pattern, spectrogram, speech intelligibility.*

Received December 7, 2017; accepted October 28, 2018
https://doi.org/10.34028/iajit/17/4/9

1. Introduction

Singing voice source segregation aims to decompose a mixture signal and extract the vocals and music from several sources in the mixture. Source segregation can be divided into two types: music segregation and speech segregation. In Speech segregation, signing voice is recovered from the mixed signal which might have some background noise as well. Whereas music segregation involves the separation of music from the mixture signal. The fact that audio signals are so intermingled and includes many variations so it is essential to segregate the desired signal from the noise. The presence of background noise in the audio signal makes it challenging for machines to effectively segregate the music and speech components. Recently, speech and music segregation has gained much importance as many existing applications for singer identification [26], music annotation [24] and lyrics recognition [23] use the information extracted from the songs. The main issues in singing source segregation are the overlapping of sources in time-frequency domains and mixture of both sources in a single channel.

The single source speech segregation can be classified in two ways: the spectrogram factorization and pitch-based inference techniques [21]. The spectrogram factorization method uses the excess of music and speech to decompose the signal into groups of repetitive sections. Each section is then allocated to a sound source. Whereas, pitch based techniques use the contour of extracted voice to separate the singing vocal. But spectrogram factorization and pitch-based techniques encounter few limitations. In [16], the repetition is said to be the foundation of music as an art. Music theorists discovered that the repetition concept is very significant in examination of the musical organization. Most of the existing methods for source segregation do not clearly count the repeating structure analysis.

Existing segregation methods can be categorized into supervised learning-based [1] and unsupervised learning-based approaches. Supervised learning-based approaches use labeled data from many sources to separate music from the speech component. On the contrary, unsupervised learning-based approaches perform source separation without having training data beforehand. Typical music/voice separation methods work by either training music model from non-verbal segments [9] or voice signal model using predominant pitch contour [10, 14]. In [3], a hybrid model is used to train both signals which require vocal segments beforehand that can be extracted using audio features such as pitch, energy, Mel-Frequency Cepstrum Coefficients (MFCCs), etc., Furthermore, other complex algorithms such as Robust Component Analysis [8] and variations of Negative Matrix Factorization (NMF) [25] can be used for monaural source separation. However, these methods are computationally expensive. In [7], non- NMF model is presented for unsupervised separation of voice in single source music signal. In speech and music segregation, different methods segregate speech from music by initially detecting voice/music segments. Classifiers such as Neural Networks (NN) and SVM identify voice and music segments using the features such as Perceptual Linear Predictive coefficients (PLP),

MFCCs, and Log Frequency Power Coefficients (LFPC). In [20], a method using NMF is proposed to segregate the spectrogram into voice and music segments [20]. But NMF requires an accurate initialization and correct number of constituents for a useful segregation. A Deep Neural Networks based technique [18] is applied to separate music and vocals from a song. In [12] SVMs are used to tag Punjabi sentences.

We suggest a different approach as compared to existing algorithms [4, 9, 10, 11, 12, 13, 14, 16, 17, 21] for source separation. The basic concept of separation is to detect all repeating periods in mixed signal and extract the repeating pattern from a non-repeating segment. The results obtained from segregation algorithm are further enhanced to improve its intelligibility.

In this paper, we segregate music and speech components by extracting repeating patterns from the mixture signal. We detect repeating structures using repeating pattern detection process and then compare these patterns with the median model. We extract the repeating segments using a time-frequency soft mask. We use an Ideal Binary Mask (IBM) filter to enhance the segregated speech and music. The use of IBM improves the affected target speech by retaining only the high energy regions of the music and speech components.

2. Proposed Methodology

We propose an effective technique to segregate the recurring background from the non-recurring foreground. The fundamental concept is to detect the patterns which repeat periodically in the song, check them with a repeating model and then extract the repeating structure. The rationale to this approach is, music signal contains a rhythm that follows a repeating pattern at regular intervals as compared to a speech signal which involves distinct variations. To improve the segregation performance, we apply an ideal binary mask on the extracted components.

The proposed technique is fast and simple as it does not create any feature vector using the acoustic features (e.g., LPC, MFCC, ZCR, etc.) to train a classifier (e.g., SVM, CNN, etc.,) for speech and music segregation. The proposed segregation technique consists of repeating patterns identification, modeling repeating sections, extracting repeating sections, and applying the ideal binary masking. The process flow diagram of the proposed system is shown in Figure 1.

2.1. Repeating Pattern Detection

Auto-correlation is commonly used to find periodicities in a signal. It computes the similarity between the current and its previous section with the passage of time. The auto-correlation method can be used to find the beat spectrum of the signal. The beat spectrum helps to compute over-all periodicity of the signal whereas, to find local periodicities we calculate the beat spectra over successive windows. Thus, beat spectrogram helps to deal with variations in periodicities with time. We apply Short Time Fourier Transform (STFT) referred as T with a hamming window of A samples using half overlap size on mixture signal y**.** We take the magnitude spectrogram S of mixture signal by keeping the absolute values of T and DC part while discarding the symmetric part. Provided the window of size $x \leq u$, where u denotes the number of time frames, we compute the beat spectrum d_t of local magnitude spectrogram S_t for each time frame t in S. After computing local beat spectrums, we combine them into beat spectra matrix D. Then we calculate the auto-correlation of magnitude spectrogram S^2 and derive matrix R_t. We use S^2 to highlight the points of periodicities in R_t**.** In case of stereo type signal, we take average of S^2 over channels. To achieve the complete sound self-similarity D of mixture signal y we take the mean of $R_t$over rows. We calculate the local magnitude spectrogram S_t, auto-correlation matrix R_t and beat spectrum d_t by using Equations (1), (2), and (3). Whereas, the beat spectra matrix D is calculated by applying Equation (4).

$$S_t(h,i) = S\left(h, i+t-\left\lceil \frac{x+1}{2} \right\rceil\right) \quad (1)$$

$$R_t(h,l) = \frac{1}{x-l+1}\sum_{i=1}^{x-l+1} S_t(h,i)^2 S_t(h,i+l-1)^2 \quad (2)$$

$$d_t(l) = \frac{1}{v}\sum_{h=1}^{v} R_t(h,l) \quad (3)$$

$$D(l,t) = d_t(l) \quad (4)$$

Where h=1…v $\left(v = \frac{A}{2} + 1\right)$ represent frequency channels, x is the window size,t shows lag and l denotes the total number of time frames.

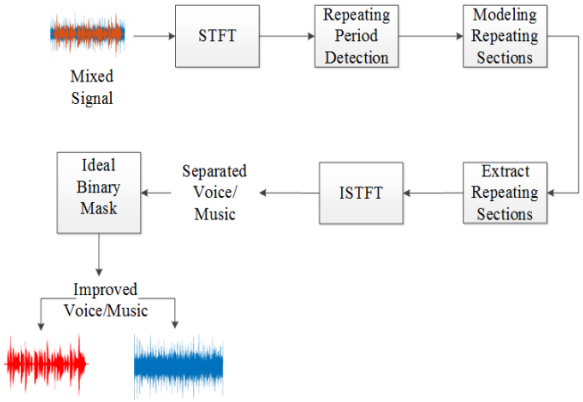

Figure 1. Process flow of proposed system.

The beat spectrum concept presented in [5] reveals that this technique allows clear understanding of the beat pattern in the mixed signal. The beat structure is referred to as the beat spectrogram. After receiving the beat spectrogram, we discard the first term which matches the overall correspondence with the signal.

Whenever there is a repeating pattern in the signal, D forms a repeating peak at different points and a peak demonstrates the repeating organization of the segment. We apply an efficient process to automatically compute the repeating pattern duration. The idea is to find the period with highest mean energy calculated among its integer multiples. To find this, we take each potential interval k in D and observe its integer multiples $j(i,e,2k,3k\ etc)$ for highest peaks in the respective neighborhood $[j-\Delta, j+\Delta]$ where Δ shows the distance and is a function of k. If its integer multiples have the highest peaks, we add their values and subtract the neighborhood's mean value to avoid the background interferences. We compute the mean energy by dividing the sum of highest peaks by all multiples of k that are in D. We refer the period P_t as the interval k which offers the highest peak values of mean so we can find the strongest peaks showing the repeating pattern of the mixed signal y. We ignore the longest lag terms to measure similarity between the segments because longest terms of autocorrelation are unreliable. To establish a repeating structure model, we need minimum of three complete repeating cycles in D. To accommodate the beat deviations we introduce a variable δ referred to as fixed variation and its length is set to 2 intervals. It means that the value of largest peak in the neighborhood is maximum of $[k-\delta, k+\delta]$. In Figure 2 we illustrate the repeating pattern of the mixed signal where each peak shows the repeating interval.

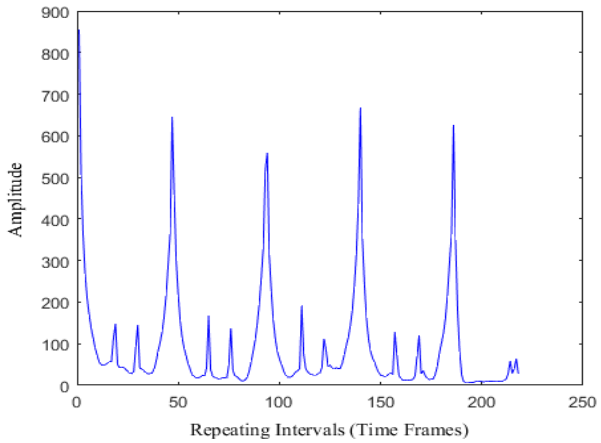

Figure 2. Repeating pattern of a mixed signal.

2.2. Modeling Repeating Sections

Once we get the repeating intervals $P_t's$, we use them to timely divide spectrogram S into g sections whose length is same as P_t. We compute median of each element of time-frequency bin at time t to obtain the repeating section and consider this median to be repeating section model O. We derive this model as:

$$O(h,t) = \underset{l=1\ldots g}{median}\left\{S\left(h,t+\left(l-\left\lceil\frac{g}{2}\right\rceil\right)P_t\right)\right\} \qquad (5)$$

Where g represents the maximum number of repeating sections and P_t indicates the length of repeating interval for frame t.

The vocal representation is different from the music in time-frequency domain and it is usually scattered. The repeating structure forms the time-frequency bins of minor variations, whereas the vocal structure that doesn't repeat periodically has larger variations. This model retains the repeating sections and discards the sections of longer variations. We refer to this model as median model [17] which proved effective to differentiate repeating sections from the non-repeating sections. We use median to get the first steady segment from the beat band because we need at least 3 stable segments to establish a median model.

2.3. Extracting Repeating Sections

After getting the repeating segment model O, we compare each segment g of the mixture spectrogram S with the repeating segment, which is the median of all segments of mixture spectrogram. Now we calculate the repeating spectrogram Q from the mixture spectrogram S. For $(S\text{-}Q)$,Q (repeating spectrogram) must be less than or equivalent to S (e.g., $Q \le S$). We compute the element-wise minimum between S and O. If the repeating segment is smaller than a segment of the mixture spectrogram, then we replace this segment with the repeating segment. The rationale is that if the value of a segment of the mixture spectrogram is larger than the repeating segment, it indicates that this segment contains more non-repeating information. In order to remove the non-repeating pattern, we need to replace this segment with the repeating segment. Otherwise, if the value of a segment is smaller than the repeating segment, it means that this segment contains less non-repeating patterns and we retain it. After comparison and replacement, the new spectrogram we derive is called the repeating spectrogram. Once we obtain the repeating spectrogram, we start to remove the non-repeating segment from the mixture spectrogram. We calculate the repeating spectrogram as follows:

$$Q(h,t) = \min\left\{O(h,t), S(h,t)\right\} \qquad (6)$$

We normalize Q by S to derive a soft time-frequency mask G using the repeating spectrogram model Q. This mask is designed on the idea that repeating time-frequency bins at some period P_t gets values closer to 1 in G. The values of non-repeating frequency bins at some period P_t tend to be closer to 0 in G. We refer to the repeating time-frequency bins as background (music) and non-repeating time-frequency bins are referred to foreground (speech). We derive soft time-frequency mask by applying Equation (7).

$$G(h,t) = \frac{Q(h,t)}{S(h,t)} \quad where \quad G(h,t) \in [0,1] \qquad (7)$$

Once this mask *G* is symmetrized, then we apply *G* to STFT of mixed signal *y*. We take ISTFT of the resultant STFT of mask *G* to obtain the music signal. We obtain the speech signal by subtracting music signal from the mixed signal ***y*.** For rest of the paper, we call it as Repeating Section Segregation Algorithm (RSSA). After performing this subtraction, we obtain the segregated music and speech components. Finally, an (IBM) is applied on these segregated components to obtain the enhanced music and speech signals.

2.4. Speech and Music Enhancement

In a cocktail party environment that comprises of various background noises, human speech communication usually degrades [2]. So, the presence of noise makes it difficult to separate the target signal. Similarly, separated speech and audio signals from the proposed method RSSA also gets some artifacts in terms of background noise. The ideal binary masking is useful as the decision making process becomes easy and the algorithm runs in a repetitive manner to estimate the sources. We apply IBM on the output of RSSA for speech enhancement.

2.4.1. Ideal Binary Mask

An IBM requires the target signal energy $S(t,f)$ and a masker signal energy $N(t,f)$ to enhance the desired speech signal. IBM is created by comparing the target signal energy of each component and the masker energy with a local Signal-to-Noise ratio criterion (LC). The noisy signal Y_k is the scaled version of target signal S_k and masker signal N_k. The noisy signal Y_k for IBM is formed as:

$$Y_k = S_k + N_k \tag{8}$$

Where Y_k represents the noisy signal which we provide as input to IBM for enhancement. LC and IBM are computed as follows:

$$LC = \log \frac{|S(t,f)|^2}{|N(t,f)|^2} \tag{9}$$

$$IBM(t,f) = \begin{cases} 1 & ,if\ S(t,f) - N(t,f) > LC \\ 0 & ,otherwise \end{cases} \tag{10}$$

In Equation (10), local criterion for the IBM is calculated by taking the log of absolute squares of target signal energy $S(t,f)$ and masker signal energy $N(t,f)$. If the difference of target energy and masker energy is greater than LC, then the time-frequency unit is assigned to target source otherwise it is set to 0. The masker signal N_k is the noise we add to target signal S_k. We take the speech/music output of RSSA as noise to form the noisy signal. The target signal S_k is referred to as a clean signal for convenience of discussion in the paper. We extract the clean signals for speech and music from the dataset as it has separate channels for each source. The input signal for IBM is referred to as noisy signal for the rest of the paper. Now the value selection for threshold LC is important because it can increase the speech intelligibility. The LC value for IBM lies in the range of 0 to -12 depending upon the application. A study [13] shows that the value of LC=-6 is better than commonly chosen value 0. For T-F analysis of IBM, both the target signal and masker is available for mixing. Discrete-time STFT is performed on the noisy and clean target signals.

3. Performance Evaluation

We used Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifacts Ratio (SAR) and Global-Normalized-Signal-to-Distortion Ratio (GNSDR) for performance evaluation.

3.1. Dataset

For our proposed system we use a standard dataset named MIR-1K [11] comprising of 1000 songs samples in WAVE format recorded at sampling rate of 16-kHz. These samples are collected from 110 Chinese karaoke songs. The voice and music is recorded separately on left and right channels respectively. We used 1000 songs clips of MIR-1K dataset to create three different mixture sets. We mix each clip of dataset into a single channel mixture using three different voice-to-music ratios of (-5, 0, 5 dB). For -5 dB music sound is louder, at 5 dB voice is louder whereas, at 0 dB song has the same effect as the original sample.

3.2. Performance Metrics

For source separation performance measurement, we use the BSS_EVAL toolbox3 [22]. It consists of a set of measures which calculate the quality of segregation among a source and its estimation. The estimate $\hat{r}$ of a source r is decomposed as follows:

$$\hat{r}(t) = r_{target}(t) + o_{interf}(t) + o_{noise}(t) + o_{artif}(t) \tag{11}$$

Where r_{target} is an acceptable distortion of source r, o_{interf}, o_{noise} and o_{artif} denote the interferences of unwanted sources, the perturbation noise, and the artifacts introduced by the separation algorithm respectively [6]. The performance measures: (SDR), Source-to-Interferences Ratio (SIR) and Source-to-Artifacts Ratio (SAR) are computed as follows:

$$SDR = 10\log 10\left(\frac{\|r_{target}\|^2}{\|o_{interf} + o_{artif}\|^2}\right) \tag{12}$$

$$SIR = 10\log 10\left(\frac{\|r_{target}\|^2}{\|o_{interf}\|^2}\right) \tag{13}$$

$$SAR = 10\log 10\left(\frac{\|r_{target} + o_{interf}\|^2}{\|o_{artif}\|^2}\right) \tag{14}$$

High values of SDR, SIR, SAR indicate better segregation of speech and music. We select these metrics as they are commonly used to evaluate the performance of segregation algorithms. These measurements are well correlated to human assessments of signal quality. Following the framework applied in [7], we calculate the Normalized Signal to Distortion Ratio (NSDR) which shows the SDR enhancement between estimate $\hat{r}$ of a source r and the mixture y as shown in Equation (15). With NSDR values we further compute the GNSDR by applying Equation (16). GNSDR represents the overall segregation performance and calculated in the proposed method by taking mean of NSDR over all mixtures set y_k weighted by their length v_k. The higher values of NSDR and GNSDR depict better segregation of speech and music signals.

$$NSDR(\hat{r}, r, y) = SDR(\hat{r}, r) - SDR(y, r) \quad (15)$$

$$GNSDR = \frac{\left(\sum_k v_k \, NSDR(\hat{r}_k, r_k, y_k)\right)}{\sum_k v_k} \quad (16)$$

3.3. Experimental Setup

We mix each song clip at voice-to-music ratio of -5, 0 and 5 dB to form a single channel signal. We compute STFT of all mixtures using half-overlapping hamming window of 40 milliseconds at 16 kHz. Then we take absolute spectrogram of the input sample. To get more precise repeating pattern, we use adaptable window and step size calculated by using Equation (1). Then we derive soft-time frequency mask by applying Equation (7). To separate voice from the music signal we also apply high-pass filter of 100 Hz cut off frequency which retain the elements of 100 Hz energy in voice segment. Whereas, transferring the energy point below 100 to music segment. The reason for selecting the cut off frequency value of 100 Hz is that the singing voice is usually above 100 Hz. For improvement of separated voice and music signals we apply the ideal binary mask on separated results. We apply IBM on all three mixtures of -5, 0 and 5 dB.

For IBM computation three types of signals are involved i.e. clean signal/target signal, noisy signal (mixture of clean signal and noise/masker) and the output/processed signal. We set the Signal-to-Noise-Ratio (SNR) of noisy speech as 0 dB and LC to -5 dB. The analysis frame duration and analysis frame shift values are set to 32ms and 4ms respectively. For IBM, the time-frequency units which have larger energy than the masker is retained while others are set to zero. This is an iterative process and continues till the target source is estimated.

3.4. Experimental Results and Discussion

We evaluated the performance of the proposed method on MIR-1K dataset. The basic idea to separate voice from music is to detect repeating intervals in the input signal. We detect repeating intervals using the auto-correlation method which finds the similarity of an interval from its previous samples. Once we compute similar segments of the signal then we use them to find stable repeating segments of the entire signal using repeating section model. We finally extract repeating spectrogram and apply soft mask on repeating spectrogram to separate music from the voice. Figures 3 and 4 illustrates the segregation results of RSSA for voice and music in terms of GNSDR, SDR, SIR and SAR. It can be observed from Figure 3 that the results for speech segregation are improved with increased voice-to-music ratio in the mixture signal. Whereas, almost the opposite trend in the results are observed for music segregation as shown in Figure 4. The reason behind this effect is that at -5 dB music has more contribution in mixture signal as compared to the speech. Since music is already dominant, so it is easy to segregate music from the speech. On the other hand, for speech segregation best results are obtained at 5 dB as speech has more proportion at 5 dB in comparison of the music signal.

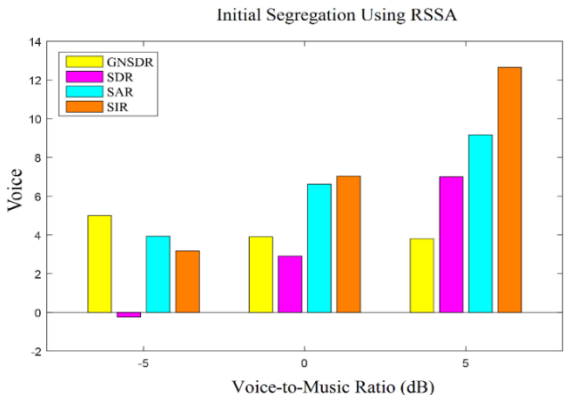

Figure 3. Segregation Results for voice using RSSA at -5, 0 and 5 dB Voice-to-Music Ratio.

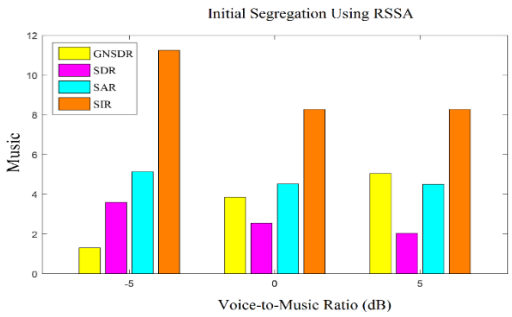

Figure 4. Segregation Results for music using RSSA at -5, 0 and 5 dB Voice-to-Music Ratio.

The segregation results shown in Figures 3 and 4 contain the noise and other artifacts that must be removed to further improve the quality of segregated signals (speech/music). For this we apply IBM on the output of RSSA in order to enhance the speech and music signals. IBM application enhances the quality of the segregated music and speech signals by removing the noise and other artifacts. IBM is applied on the output of RSSA for speech and music signals and results

are presented in Figures 5 and 6.

Shown in Figures 5 and 6 are the enhanced results in terms of GNSDR only, as this is the average value of segregation for complete dataset. Hence provides better measure of segregation for the entire data. IBM compares noisy signal energy with the clean signal energy based upon a local criterion as discussed in the methodology section, whereas retaining high target energy parts and discarding the rest. It can be observed from Figure 5 that combining RSSA with the IBM provides better results for both the music and speech signals. Moreover, IBM performs best at 5 dB due to high speech-to-music ratio.

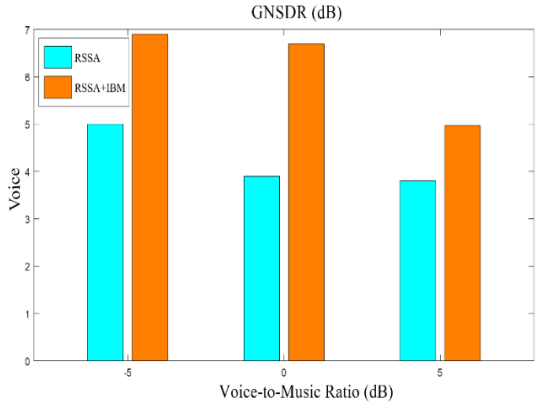

Figure 5. Segregation Results of RSSA and RSSA+ IBM for Voice -5, 0 and 5 dB Voice-to-Music Ratio.

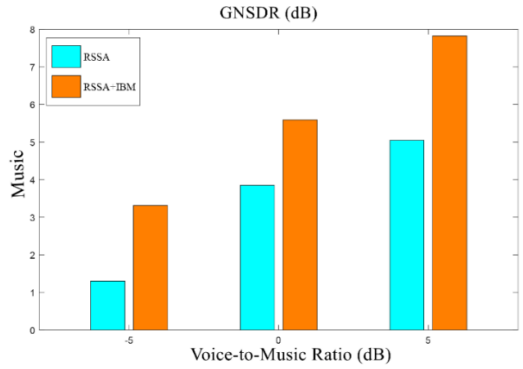

Figure. 6. Segregation Results of RSSA and RSSA+ IBM for Music -5, 0 and 5 dB Voice-to-Music Ratio.

To further elaborate on the performance of IBM, we show amplitude waveforms and frequency spectrograms (Figures 7 and 8). The first column shows the amplitude waveform at different stages of the algorithm (i.e., target signal, noisy signal and processed signal) from top to bottom respectively. The second column shows the frequency spectrograms of respective stages.

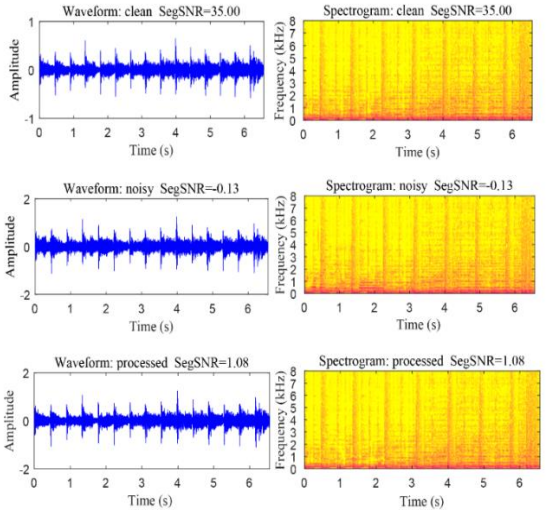

Figure 7. IBM results for Music.

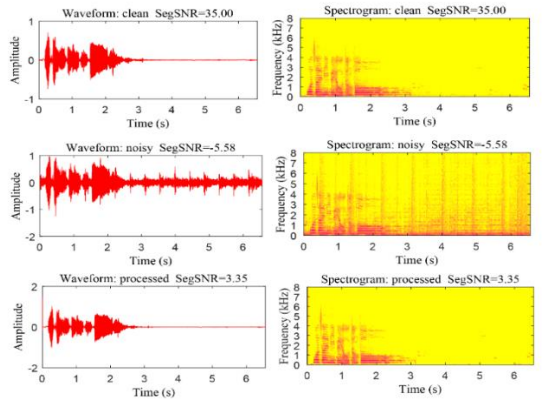

Figure 8. IBM results for Speech.

It is evident from Figures 7 and 8, that in case of music signal, amplitude waveform and frequency spectrogram is nearly same for both target and noisy music signal. The SNR value of music and speech signal is negative after the application of RSSA. The value of SNR represents the quality of a signal, where positive value indicates noise free signal and negative value indicates the noisy signal. After applying IBM, quality of the signal improves as IBM retains high energy of target signal while discarding low energy components. High energy components are more likely to be a target signal compared to low energy components. An important factor in the speech intelligibility is the right choice of LC. We check the range of LC's between -6 dB to 6 dB. In our case the values greater than LC=-5 yield poor results. Whereas, LC=-6 provides better results in terms of SNR again but LC=-5 provides the best results over all three mixture set (-5, 0 and -5 dB). The application of IBM in combination of RSSA provides better segregation performance in terms of GNSDR.

Shown in Figure 9 is the performance comparison of the proposed method against existing state-of-the-art methods for music and speech segregation. The proposed method outperforms state-of-the-art methods at -5 dB and 5 dB mixtures while generating comparable results at 0 dB. The results are only compared for speech

segregation as majority of the existing methods have presented the results of speech segregation only. We report the comparative analysis using GNSDR value for segregated speech signal as it provides the performance measure on the entire dataset. The performance of existing segregation methods [10] degrades significantly on GNSDR value at-5 dB because their technique performs well on higher voice-to-music ratio. Note that the proposed method achieves highest GNSDR values at -5 dB because music contribution is higher at-5 dB mixture and RSSA extract the repeating patterns effectively. In addition, IBM enhances the segregated output which ultimately increases the GNSDR values. The proposed method achieves GNSDR values of 12.91, 8.67 and 7.97 at -5 dB, 0 dB and 5 dB respectively. The proposed method achieves reasonably good efficiency with better segregation results by combining the RSSA with improved binary masking approach.

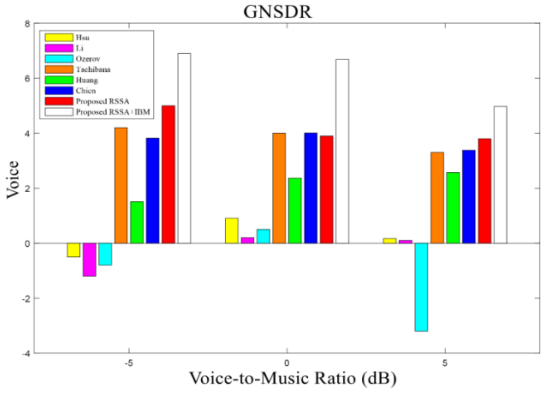

Figure 9. Comparison of proposed method with Hsu [10], Li [14], Ozerov[15], Tachibana [19], Huang[8], Ikemiya [11],Tzung Chien [1].

References

[1] Chien J. and Yang P., "Bayesian Factorization and Learning for Monaural Source Separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 185-195, 2015.

[2] Colin Ch., "Some Experiments on The Recognition of Speech, with one and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975-979, 1953.

[3] Durrieu J., David B., and Richard G., "A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1180-1191, 2011.

[4] Deif H., Fitzgerald D., Wang W., and Gan L., "Separation of Vocals from Monaural Music Recordings Using Diagonal Median Filters and Practical Time-Frequency Parameters," *in Proceedings of IEEE International Symposium on Signal Processing and Information Technology*, Abu Dhabi, pp. 163-167, 2015.

[5] Foote J. and Uchihashi S., "The Beat Spectrum: A New Approach to Rhythm Analysis," *in Proceedings of IEEE International Conference on Multimedia and Expo*, Tokyo, 2001.

[6] Févotte C., Gribonval R., and Vincent E., "BSS_EVAL Toolbox User Guide--Revision 2.0," HAL-Inria, pp. 1-19, 2005.

[7] Hu Y., Wang L., Huang H., and Zhou G., "Monaural singing voice separation by non-Negative Matrix Partial Co-Factorization with Temporal Continuity and Sparsity Criteria," *in Proceedings of International Conference on Intelligent Computing*, Lanzhou, pp. 33-43, 2016.

[8] Huang P., Chen S., Smaragdis P., and Hasegawa-Johnson M., "Singing-Voice Separation from Monaural Recordings Using Robust Principal Component Analysis," *in Proceedings of Acoustics, Speech and Signal Processing, IEEE International Conference On*, Kyoto, pp. 57-60, 2012.

[9] Han J. and Chen C., "Improving Melody Extraction Using Probabilistic Latent Component Analysis," *in Proceedings of IEEE International Conference on Acoustics Speech and Signal*, Prague, pp. 33-36, 2011.

[10] Hsu C. and Jang J., "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *IEEE Transactions on Audio, Speech, and Language* vol. 18, no. 2, pp. 310-319, 2010.

[11] Ikemiya Y., Itoyama K., and Yoshii K., "Singing Voice Separation and Vocal F0 Estimation Based on Mutual Combination of Robust Principal Component Analysis and Subharmonic Summation," *IEEE ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2084-2095, 2016.

[12] Kashyap D. and Josan G., "Prediction of Part of Speech Tags for Punjabi Using Support Vector Machines," *The International Arab Journal of Information Technology*, vol. 13, no. 6, pp. 603-608, 2016.

[13] Li Y. and Wang D., "On The Optimality of Ideal Binary Time-Frequency Masks," *Speech Communication*, vol. 51, no. 3, pp. 230-239, 2009.

[14] Li Y. and Wang D., "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1475-1487, 2007.

[15] Ozerov A., Philippe P., Bimbot F., and Gribonval R., "Adaptation of Bayesian Models for Single-Channel Source Separation and Its Application To Voice/Music Separation In Popular Songs," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1564-1578, 2007.

[16] Rafii Z. and Pardo B., "Repeating Pattern Extraction Technique: A Simple Method For Music/Voice Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 1, pp. 73-84, 2013.
[17] Schenker H., Harmony EM Borgese, Cambridge University Press, 1954.
[18] Sharma V., "A Deep Neural Network Based Approach for Vocal Extraction From Songs," *in Proceeding of Signal and Image Processing Applications IEEE International Conference on*, Kuala Lumpur, pp. 116-121, 2015.
[19] Tachibana H., Ono N., and Sagayama S., "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, pp. 228-237, 2014.
[20] Vembu S. and Baumann S., "Separation of Vocals From Polyphonic Audio Recordings," *in Proceeding of International Society for Music Information Retrieval*, London, pp. 337-344, 2005.
[21] Virtanen T., "Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066-1074, 2007.
[22] Vincent Emmanuel., http://bass-db.gforge.inria.fr/bss_eval/Last Visited, 2018.
[23] Wang C., Lyu R., and Chiang Y., "An Automatic Singing Transcription System with Multilingual Singing Lyric Recognizer and Robust Melody Tracker," *in Proceeding of 8th European Conference on Speech Communication and Technology*, Geneva, pp. 1197-1200, 2003.
[24] Yang D. and Lee W., "Disambiguating Music Emotion Using Software Agents," *International Conference on Music Information Retrieval*, vol. 4, pp. 218-223, 2004.
[25] Zhu B., Li W., Li R., and Xue X., "Multi-Stage Non-Negative Matrix Factorization For Monaural Singing Voice Separation," *IEEE Transactions on Audio, Speech, And Language Processing*, vol. 21, no. 10, pp. 2096-2107, 2013.
[26] Zhang T. and Packard H., "System and Method for Automatic Singer Identification," *Research Disclosure*, pp. 756-756, 2003.

Sidra Sajid received her M.Sc. degree in Software Engineering from UET Taxila, Pakistan in 2018. She is working as IT Officer in Primary and Secondary Healthcare Department, Punjab. Her research interests include audio signal processing, classification problems and machine learning.

Ali Javed received his Ph.D. degree in Computer Engineering from UET Taxila, Pakistan in 2016. Currently, he is serving as Assistant Professor in Software Engineering Department at UET Taxila, Pakistan. His areas of interest are Digital Image Processing, Computer vision and Machine Learning.

Aun Irtaza has completed his PhD from FAST-National University of Computers & Emerging Sciences in 2016. Currently he is serving as HOD in Computer Science Department at UET Taxila. His research interests include computer vision, pattern analysis, and big data analytics

.