Sabaragamuwa University Journal of Computer Science

Journal homepage: https://www.sab.ac.lk/sujcs/

Department of Computing and Information Systems, Faculty of Applied Sciences, Sabaragamuwa University of Sri Lanka

Digital signatures for singers to identify their songs

H.M.S.R.Heenkenda[1],D.D.Karunaratna[2],S.M.K.D.Arunatilake[3]
University of Colombo School of Computing, Sri Lanka

Abstract

Counterfeiting is the process of imitating the voice of a popular artist, with the intention of selling or passing the imitations as genuine. This research is aimed at identifying the imitated songs by generating unique digital signatures for each singer by using songs sung by the artists.

Songs typically contain vocal signals surrounded with instrumental signals. In order to generate signatures for the voice of a singer, the vocals have to be isolated. Voice isolation and Artist classification had been addressed as two different problems throughout the past decades. Our research combines these two problems, to build a unique signature model for each singer. This study proposes a technique to isolate vocal signals from instrumental signals by using REPET filter, Harmonic - percussive source separation, Butterworth band-pass filter and silence removal. This technique show better results than the prevailing techniques for vocal isolation. The signatures are then generated by extracting features from the isolated vocal signals and represented as GMM Models. The results of this research are evaluated through quantitative and qualitative approaches by using a sample of songs sung by Sri Lankan artists. The outcome of this research has demonstrated the possibility of generating digital signatures for singers by using their songs. The technique proposed has the ability to distinguish singers having similar voices accurately. Finally, the technique proposed in this paper could be used to generate unique signatures for singers who sung similar songs.

Keywords*:* Audio Signal Processing, Gaussian Mixure Model, Harmonic Percussive source separation, Repeating Pattern Extraction Technique, Voice isolation

Received 14 May 2021
Accepted 31 July 2021

Sabaragamuwa University Journal of Computer Science

ISSN : 2783-8846

1 ruwanariheenkenda@gmail.com

2 mailto:ddk@ucsc.cmb.ac.lk

3 mailto:sda@ucsc.cmb.ac.lk

1 INTRODUCTION

In contrast to real properties, Intellectual properties are not tangible as they originate in human minds involving human intellect. Intellectual property rights in a country are the laws that make the provisions for protecting the intellectual creations of humans. A famous singer has a right to protect the use of his/her voice in order to earn an income from his/her songs. These rights are typically abused when widely known singers distinctive voices are deliberately imitated to generate and sell songs as disguises to the original singers [1]. A counterfeit is an imitation, usually made with the intention of passing it off as a genuine. Only the copyright holder of a song has the right to make commercial usage from his work. If another party tries to counterfeit the original work and make commercial usage out of it, that would come under the copyright infringement and breaking of intellectual property rights of that artist. Counterfeit product and an original work of an artist (song) will be very hard to distinguish when heard by a normal person. It requires extensive knowledge and practice in music for a typical person to distinguish between original singing and well-imitated copies.

There are a number of audio signal processing based applications used in the music industry, for applications like content-based audio identifications (CBIV), Content-based integrity verification (CBID) and watermarking support [2]. Although there are applications to provide audio fingerprints of songs, there does not exist an application to give singers a unique signature to their voice by giving their songs as an input. With the capabilities of Music Information Retrieval and signal processing techniques, it is worthwhile to venture into integrating these technologies into generating a digital signature for singers to recognize their songs.

This paper describes a solution for the artists who had been affected by the dissenters who imitate their voices in song productions and performances. The technique proposed shows how digital signatures of singers can be used to distinguish between this real and fake singing. The digital signatures of singers are generated by extracting the unique features from their songs and accumulated them into a model. Subsequently, these models are verified by comparing the models with each other and classifying known songs to singers. The signature generation is done by following a process of voice isolation and feature extraction. Voice isolation has been done using a combined strategy of pre-defined voice isolation approaches.

2 LITERATURE REVIEW

Generating digital signatures for singers to identify their songs using songs as input can be considered as a novel research. Even though there had been many approaches for artist classification, no researcher had used songs as inputs. In the past, separate research had been done on voice isolation and artist identification. This following section gives a brief review on the previous methodologies used for voice isolation and artist identification.

2.1 Voice Isolation

Acoustically all kinds of sounds are similar but they possess fundamental differences. The main attribute that distinguishes musical instruments from one another is the timbre. Timbre is the quality of sound that differentiates different sounds [3]. The timbre of a sound depends on its waveform, their frequencies, and their relative intensities. Brightness and roughness, can also be helpful to understand the dimensions of timbre. It can be separated by using the expression attributes. The most common methodologies to extract features related to the timbre of sound are the Mel Frequency Cepstral Coefficients (MFCC) and Formant Analysis [4] [5]

In order to isolate the voice of a song, features of the song spectrum should be extracted. There are basically two types of approaches to extract features of sound, MFCC and Formant analysis. The MFCC are considered as a collective representation of the short-term power spectrum of a sound. This power spectrum is supported on a linear cosine transform of a log power spectrum on a frequency which is not linear. MFCC collectively create a Mel Frequency Spectrum. They are derived from a type of cepstral representation of the sound tracks (audio clips) [6]. MFCCs are commonly used as features in speech recognition systems, Genre classification and in audio similarity measures. Formants are the spectral peaks of the acoustic spectrum energy around a particular frequency in the speech wave. They can be considered as the distinguishing or meaningful frequency components of human speech and of singing.

Most researchers had used pitch estimation for voice and instrumental separation. [8] [10]. This had been done by assuming the fundamental frequency of the song lies on the vocals of the singer. The main reason why it is not used in this project is that the in our experiments we have observed that the fundamental frequency of Sri Lankan songs do not lie fully on the vocal partition of a song. The fundamental frequency of Sri Lankan songs tends to lie on the instrumental partition also. Therefore, using pitch estimation to isolate vocals had been discarded in this project.

Another popular mechanism used to isolate vocals had been the spectrogram analysis. In the

spectrogram analysis spectral discontinuity thresholds and robust component analysis had been used on detecting changes in the spectrogram. [9] [11] Robust component analysis work has primarily applied on rock and pop type songs. But most of the Sri Lankan songs has classified as in the classical genre.

Another interesting principle in songs, that had been analyzed is repeating patterns of the song and considering that partition as background music. [12] This method had been used in this research as one filter for voice isolation.

Pitch estimation has advantages and disadvantages when used for voice isolation. The main reason why it is not used in this project is that the fundamental frequency of a Sri Lankan song does not lie fully on the vocal partition of a song. It may lie on the instrumental partition too. Therefore, using pitch estimation to isolate vocals had been discarded in this project. Robust component analysis had worked primarily on songs which were of the genre rock and pop. But most of the Sri Lankan songs had been of the classical genre. Therefore, using robust component analysis had also been avoided.

2.2 Artist Identification

The K-means clustering had been the most common method used by the past research for the artist identification purposes. [13] [14] K-means is considered to be very effective in fine-tuning cluster borders locally. But it fails to relocate the centroids globally. K-means cannot remove centroids that are not needed nor can insert new centroids when there are stable clusters [7]. Therefore, using K-means for clustering is avoided in this project.

Most of the features extracted are timbral features as the vocals show their uniqueness through the primary features which are spectral energy, frequency, and timbral features.

Some researchers had used Support Vector Machines (SVM) and Gaussian Mixture Models (GMM) for artist classification. [15] [16] the next most popular methodology that had been using is the training of two different models for vocal and nonvocal sections and applying these models afterward. [17]. One of the major requirement of this approach is the availability of a large number of separated vocals of singers to train the vocal models. Due to the in-feasibility to locate that many numbers of separated vocals of singers to train the vocal models, this method has also discarded.

3 MATERIALS AND METHODS

This research is different to the similar research reported in the literature in a number of aspects. Firstly, in the previous work singer identification was done by using the audio clips containing only the vocal segments of the artist. In this research singer identification using the songs of the artists with music as the input. Secondly, this research proposes a technique to separate vocals of singers from the background music. Lastly, a digital model is proposed to record the vocal attributes of singers. The following Figure I depicts the flow of this research's approach primarily. The process to achieve each and every step in this diagram will be briefed in this section.

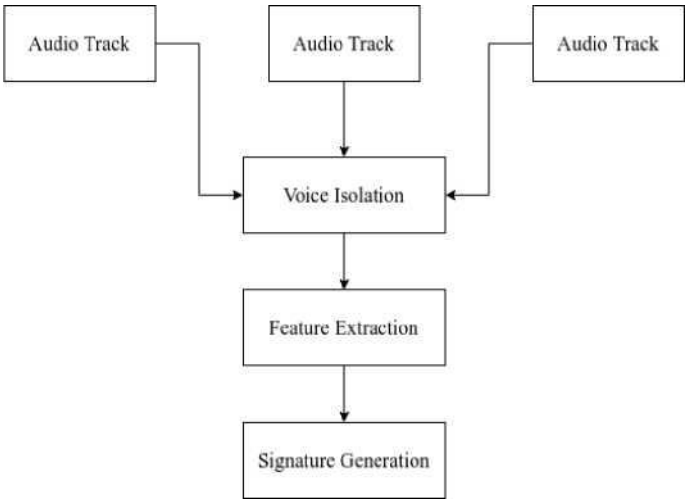

Fig. I. Primary research approach

This project is implemented using open source software. The code is written in Python language using the spyder IDE editor. It uses the Python libraries Librosa, MatplotLib, Pandas to achieve the project objectives. The proposed approaches were aimed at only Sinhala songs.

3.1 Voice Isolation

The Sinhala song use for this research consists of background music consists of sounds of various musical instruments. In this research, the sri Lankan song has been examined thoroughly and how the features of the vocals and music differ from each other was studied. In this study the following observations were made,

1) The song can be divided into harmonic and percussive partitions. Vocals are inside the harmonic partition.
2) The background music follow the same pattern. (Guitars, piano chords, drums etc.)

3) The vocals are within the frequency range 85 to 855 Hz.
4) The introductory and ending parts of songs mostly consist of instrumental music or silence.

Harmonic sounds are the ones which we observed to have a certain pitch. Percussive sounds do not have a pitch but a clear localization in time. Mostly singing vocals have harmonic features [18]. Our experiments show that the percussive sounds could be used to separate harmonic-percussive source.

The repeating pattern of a song in the instrumentals comprises of the same melody in most of the songs. Most Sri Lankan songs do not exhibit a sudden change in rhythm or melodiousness. The chord progression gives the song a particular color. In order to maintain that specific key or the color of the song, the chord progression, and the rhythm is kept the same throughout the song. Therefore, instrumental music can be unmasked by detecting a similar pattern of the song as depicted in the spectrogram. Afterward, the amplitude of the frames in which has repetition is lowered to enhance vocals.

Women typically sing in three groups of voice ranges: soprano, mezzo-soprano, and contralto. Men are typically separated into four groups: countertenor, tenor, baritone, and bass. Men's voices are deeper than ladies' as their vocal chords are longer. This classification depicts that the highest frequency range of humans as Soprano while the lowest as the Bass. The soprano's vocal range (utilizing logical pitch documentation) is considered from around middle C (C4) = 261 Hz to ''high A" (A5) = 880 Hz in choral music. The frequency range of a typical adult bass singer is said to be from 85 to 180 Hz. This concludes that the singing vocal frequency range can be defined as 85 Hz to 880 Hz.

The removal of instrumental music from a song results in a number of long periods of silence in the track. Lots of research had been conducted to show how the removal of silenced and unvoiced segments had improved the efficiency of the performance of the system. Silence removal had been very helpful portion of the proposed technique to reduce processing time and increase the performance of system by eliminating unvoiced segments from the input signal. [19] Research had been done to identify unvoiced and silenced signals to enhance the performance of the speech/vocal signal processing. Experiments were done to evaluate whether the removal of unvoiced segments has enhances the voice isolation process or not.

By using those observations, some efforts had been made to reduce the effect of the background music and enhance the vocal part of the singer. The four filters that strengthen the vocal part of singers are,

1) Harmonic Percussive source separation using median filtering.
2) Voice extraction using Similarity Matrix.
3) Butterworth Band-Pass Filter.
4) Eliminating introductory part.

3.1.1 ***Harmonic Percussive source separation:*** The technique used includes the usage of median filtering on a spectrogram of the sound signal, with median filtering performed across progressive frames to stifle percussive occasions and improve harmonic partitions, while median across frequency bins is used to strengthen percussive occasions and supress consonant segments. The two emerging median filtered spectrograms are used to generate masks which are then applied to the main spectrogram to isolate the harmonic and percussive pieces of the signal.

In the research the harmonic occasions are approximated as vertical lines and percussive occasions are approximated as horizontal lines in a spectrogram. Median filters work by replacing a given sample in a signal by the median of the sign values in a window around the example. Given an input vector x(n) and then y(n) is the yield of a median filter of length l, where l characterizes the quantity of samples over which median filtering happens. Where l is odd, the middle channel can be characterized by the following equation 1.

$$y(n) = median * (x(n-k : n + k), k = (l-1)/2) \quad (1)$$

In the recorded research there are many observed issues. One issue is that the computed components are frequently not of purely harmonic or percussive in nature yet in addition contain commotion like sounds that are neither clearly harmonic nor percussive. Besides, depending on the parameter settings, one often can watch a spillage of harmonic sounds into the percussive segment and a spillage of percussive sounds into the harmonic segment. Subsequently the methodology was extended by utilizing two expansions to a state-of-the- art harmonic-percussive separation procedure to focus on the issues. Initially, a partition factor parameter is introduced into the disintegration procedure that permits for fixing separation results and for upholding the segments to be unmistakably harmonic or percussive. The other extension was inspired from the classical sines+transients+noise (STN) sound model. This novel idea is exploited to highlight a third residual segment to the decomposition which catches

the sounds that stay between the distinctly harmonic and percussive sounds of the audio signal.

The Figure II shows how the spectrogram of a song looks like before and after performing the harmonic-percussive source separation.

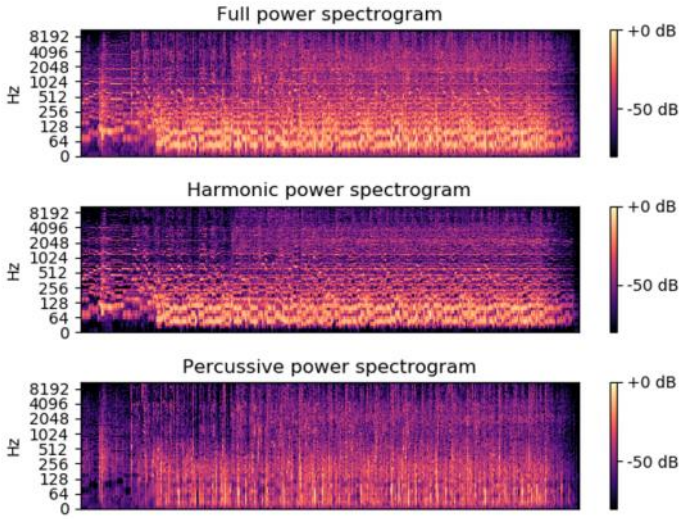

Fig. II. Harmonic-percussive source separation

3.1.2 ***Voice Extraction using Similarity Matrix****:* A similarity matrix is defined as two-dimensional representation where each point (a, b) measures the dissimilarity between any two elements a and b of a given sequence. Since, repetition is mostly used in the instrumental parts of Sinhala songs, a similarity matrix calculated from an audio signal aids to discover the musical structure that underlies it. Given a single-channel mixture signal x, first, its Short-Time Fourier Transform would be calculated using half overlapping Hamming windows of a particular length. Then the magnitude spectrogram V is derived by taking the absolute value of the elements of X, after discarding the symmetric part, while keeping the direct current component. The similarity matrix S is then defined as the matrix multiplication between transposed V and V, after normalization of the columns of V by their Euclidean norm.

After the similarity matrix is calculated, the repeating elements can be found in the mixture spectrogram. For all the frames j in mixture spectrogram, the frames which are the most close to the given frame are identified and saved as a vector.

By following the rationale, “the non-repeating foreground (voice) has a sparse time-frequency representation compare to the time-frequency representation of the repeating background (music)”, the researches had come to a conclusion that time frequency bins with little deviations between repeating frames would constitute a repeating pattern and would be

captured by the median. Once the repeating elements have been identified for all the frames j in the mixture spectrogram V through their corresponding vectors of indices, they had been used to derive a repeating spectrogram model W for the background by taking the median of the k number of frames.

After generating the repeating spectrogram model, a time frequency mask is derived by normalizing the repeating spectrogram model. The time-frequency mask is then symmetrized and applied to the Short time Fourier transform of the mix-ture signal x. The estimated music signal is finally obtained by inverting the resulting STFT into the time domain. The estimated voice signal is obtained by simply subtracting the music signal from the mixture signal. This filter is defined as the Repeating Pattern Extraction Technique (REPET) [12] The Figure III shows how the spectrogram of a song looks like before and after using the REPET filter.

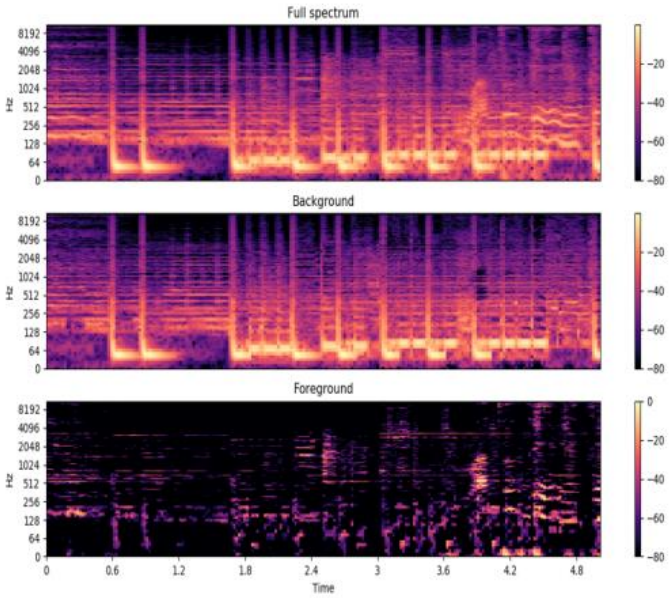

Fig. III. Isolation of vocals using REPET

3.1.3 Frequency Filtering using a Band-pass filter: The Butterworth filter is expressed as a form of signal processing filter designed to have a frequency response as flat as achievable in the passband. It is additionally referred to as a maximally flat magnitude filter. The frequency response of the Butterworth filter is maximally flat within the passband and rolls off approaching zero in the stopband. [20] When viewed on a logarithmic plot, the response is a slope which declines off linearly towards negative infinity. A first order filter's response falls off at −6 dB per octave (−20 dB per decade) (all first-order lowpass filters have identical normalized frequency response). A second-order filter decreases at −12 dB per octave, a third-order at −18 dB and likewise. Butterworth filters have a monotonically wavering magnitude

function with ω, unlike other filter forms that have non-monotonic ripple in the passband and/or the stopband.

This filter can be compared with a Chebyshev Type I or the Chebyshev Type II filter or an elliptic filter. From all those, the Butterworth filter has a very slow roll-off. That is the reason why it requires a higher order to implement a specific stopband specification. But Butterworth filters have an extra linear phase response in the pass-band than the mentioned filters can accomplish. The Figure IV shows how the spectrogram of a song looks like before and after using the band-pass filter.

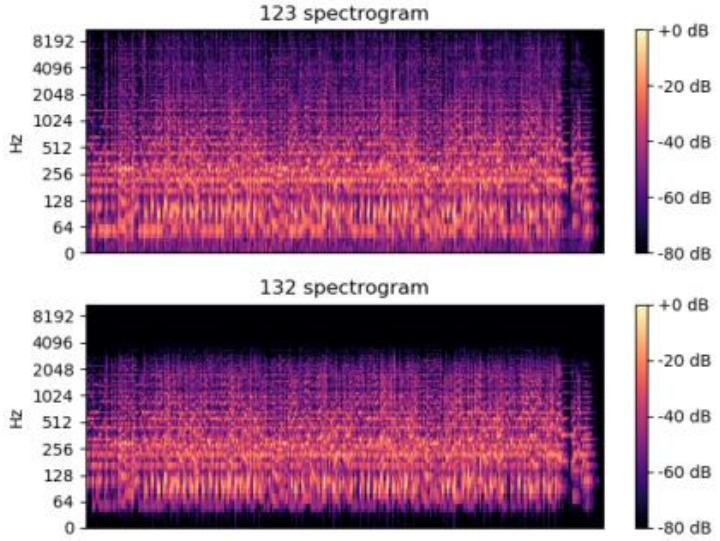

Fig. IV. Frequency filtering using band-pass filter

3.1.4 Eliminating introductory and ending parts: A survey analysis has been conducted to observe the significant amount of time used in the introduction and the end of the song consisting with only instrumental music or silence in the case where after vocals were being isolated. As sound is a vibration, there is a capability to access frame by frame and to check whether it is silent or not. This has been achieved in this research using the PyDub library in python. The silence threshold used in this project has been −50 decibels. The threshold had been found through a trial and error methodology, as per the audios have been of the same quality. The reasoning behind the usage of trial and error method for this functionality is the silence threshold depends hugely on the quality of the audio and the duration of the silence in the audio. When listening to the audio after trimming the introduction, the audio had been in a satisfactory level in this project.

3.2 Feature Extraction

The audio signal is a three-dimensional signal in which represent time, amplitude and frequency. The features suitable for speech signals were selected for this project. The features chosen are, MFCC, zero Crossings Rate, Spectral Centroid and Spectral Rolloff.

3.2.1 MFCC: Mel-Frequency Cepstral Coefficients (MFCCs) is defined as a formation of dimensionality reduction. One might pass a collection of audio samples, and receive 10 to 20 cepstral coefficients that describes that sound in a typical MFCC computation. While MFCCs were initially developed to represent the sounds made by the human vocal tract, they have turned out to be a pretty solid timbral, pitch invariant feature, that has all sorts of uses other than automatic speech recognition tasks. When obtaining MFCCs, the first step is to compute the Fourier transform of the audio data, which converts time domain signal into a frequency domain signal.

3.2.2 Zero-Crossings Rate: The zero-crossing rate is the rate of the sign being changed along a signal, the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval whereas it usually has higher values for highly percussive sounds like those in metal and rock. Therefore the artists are classified according to the significant features of their voices for and example, a rock singer usually has higher zero crossings rate when compared with a classical singer. It had seem that artist classification can be achieved through this feature and hence this feature was extracted from each and every isolated vocal.

3.2.3 Spectral Centroid: This depicts the place where the "center of mass" for a sound is located. It is calculated as the weighted mean of all the frequencies present in the sound. If the frequencies in music are similar or around a similar frequency throughout the whole song, then spectral centroid would be around a center and if there are high frequencies at the end of sound then the centroid would be mapped towards its end. Spectral centroid is considered a good indicator of brightness. In music, timbre which is also known as tone color, is considered the recognized sound quality of a musical note, sound or note. Timbre is the term referred to as brightness here. Timbre distinguishes different types of vocals. As the spectral centroid can distinguish between the tonal colour or the timbre of vocals this feature had to be extracted. [22]

3.2.4 Spectral Rolloff: Spectral rolloff is considered the frequency below which a specified major percentage of the total spectral energy, e.g. 85 percent, lies.It also gives results for each frame. Kos et.al. [23] had discussed the usage of spectral roll-off in acoustic classification and emphasized music/voice classification in their paper The spectral roll-off is a timbre feature. As it produces features of the timbre of voice, that feature had been extracted and used in this research project.

3.3 Signature Generation

Saini et.al. have discussed in the research paper [21] how using MFCC features with the GMM model had given an exceptional performance in most of the speaker identification tasks. They had also mentioned as Speaker recognition is more of a biometric task, speaker-related recognition activities like artist recognition can presumably have a better ending when a Gaussian Mixure Model is used. Therefore a Gaussian Mixture model had been used in the final process: signature generation.

A Gaussian mixture model is considered as a probabilistic clustering model to represent the presence of sub populations within an overall population. The reason behind raining a GMM is to approximate the probability distribution of a class by a linear combination of ‘k' Gaussian distributions. The likelihood of feature vectors for a model is given by following equation 2:

$$P(X/\lambda) = \sum_{k=1}^{K} \omega_k P_k(X/\mu_k \sum k) \quad (2)$$

, where $P_k(X/\mu_k \sum k)$ is the Gaussian distribution.

The training data Xj of the class A are used to estimate the parameters mean co-variance matrices S and weights w of these k components. Initially, it has identified k clusters in the data by the K-means algorithm and has assigned equal weight w = 1/k to each cluster. ‘k' Gaussian distributions are then had been fitted to these k clusters. The parameters a and w of all the clusters are updated in iterations until the converge. The mostly used method for this is the Expectation Maximization (EM) algorithm. Therefore, it can be concluded that when feature vectors unique to each singer are provided, a GMM model unique to each of the singer can be retrieved. Let X be a time series

of feature vectors selected and A be the GMM for the singer s. Then, the signature of the singer is determined through the following equation 3,

$$S =_i 1/T \sum_{t=1}^{T} logp(x_t/\lambda_i) \quad (3)$$

The signatures for each and every singer which resembles the equation mentioned above are generated using audio tracks of that artist. This signature can be considered as a unique model of that particular artist.

4 RESULTS AND DISCUSSION

The evaluation design of this research mainly focuses on the effectiveness of the pre-processing stages used in the voice isolation section, and the ability of the introduced unique signature to differentiate between presumed artists and artistes. An effective and efficient procedure is followed in order to preserve the quality and value of the research.

Many pre-processing steps had been used in the stage of voice isolation in this research namely, voice extraction using a similarity matrix, harmonic percussive source separation, frequency filtering using a Band-pass filter and the elimination of introduction and the end. REPET (Repeating Pattern Extraction Technique) introduced by Rafii and Pardo is kept as the base and the other filters are tested out combined, and separately to see their effectiveness in the voice isolation process.

The following 4 cases are the evaluation criteria we have tested in order to find the best voice isolation approach which generates the best digital signature. These evaluation tasks are carried out and the accuracy of artist identification is calculated for each case.

- Case 1: When using REPET alone for voice isolation.
- Case 2: When using REPET + Harmonic-Percussive source separation for voice isolation.
- Case 3: When using REPET + Band-pass filter for voice isolation.
- Case 4: When using REPET + Harmonic-Percussive source separation + Band-pass filter

for voice isolation.

The first case is conducted in order to see the progress when using a state-of-art method in this research. The second and third cases are conducted in order to clarify if each filter separately improves the evaluation or not. The whole process from voice isolation to signature generation is carried out and finally, the accuracy of the artists is identified as a percentage and those accuracy percentages are produced for each and every evaluation task. The evaluation task which yields the highest accuracy or in other words, the winner of these subtasks is recognized as the most suited voice isolation process to be used in this research. The accuracy for this task is calculated using the following equation 4:

$$\text{Accuracy} = \frac{\text{No: of songs correctly identified its artist}}{\text{No: of songs in the test set}} \qquad (4)$$

The winner from those four cases will be evaluated once again against the tracks where the silenced partitions of introduction and ending are eliminated. This is done in order to see if there is an improvement in eliminating silence and unvoiced partitions of the song.

4.1 Data

The dataset consisted of approximately 600 songs of 14 female and 16 male artists having 20 songs per each artist. The training and testing data were split from this dataset, training set having 70% of the data while test set having 30% of the data. The data had been randomly chosen to be included in test or training sets. This research approach uses for monaural (single channel) songs in the mp3 (MPEG-1 standard) format as input and produces the digital signature of the corresponding singer as the output.

4.2 Results

4.2.1 REPET alone: The dataset for this evaluation criteria has been trained only using the REPET filter. GMM models are generated for every singer and the accuracy is obtained as the percentage of test data which were correctly identified the label of the model as the appropriate

singer. The accuracy of the signature generation when using REPET alone had been 0.4166667.

4.2.2 REPET + Harmonic Percussive source separation: Here, the data set had been preprocessed using both harmonic percussive source separation and REPET together. The accuracy of signature generation had been 0.61111.

4.2.3REPET + Band-pass filter*:* Both REPET and the butterworth band-pass filter had been used in the preprocessing stage of this evaluation. The accuracy obtained is 0.53448.

4.2.4REPET + Harmonic percussive separation + Bandpass filter*:* All three filters had been combined in this experiment. All the filters had been performed on the songs in training and test datasets. The result had been the best as of then, resulting in 0.743589 accuracy.

4.2.5Discussion of combined approaches*:* These results have been represented as rounded off percentages in the Table I.

TABLE I

ACCURACIES OF THE FILTER COMBINATIONS

#	Method	Accuracy
1	REPET	42%
2	REPET + HP	62%
3	REPET + BP	54%
4	REPET + HP + BP	74%

The performance of voice isolation for the task of generating digital signatures for singers had been improved by using the harmonic percussive source separation and band-pass filter significantly. These results show that the best approach for voice isolation is the combination of all three REPET, Harmonic percussive source separation and using the Band-pass filter. These percentages show some interesting theories like, harmonic percussive source separation being more convenient than the band-pass filter for this particular research question.

4.3 Effect of silence removal

The winning approach is again evaluated against the winning approach combined with silence removal. This filter had been used finally because, silence removal filter works well with completely isolated vocals (instrumental music completely removed). Therefore the best voice isolation approach is first identified and then the silence removal is evaluated. The winning voice isolation approach is declared as the combination of all three filters. Therefore, after applying all REPET, HP and BP filters, silence of introduction and ending are removed and evaluated. The resulted accuracy of this combination was 0.717948. The Figure V depicts how the silence removal's effectiveness was evaluated against the winning voice isolation combined approach (REPET + HP separation + Bandpass filter)

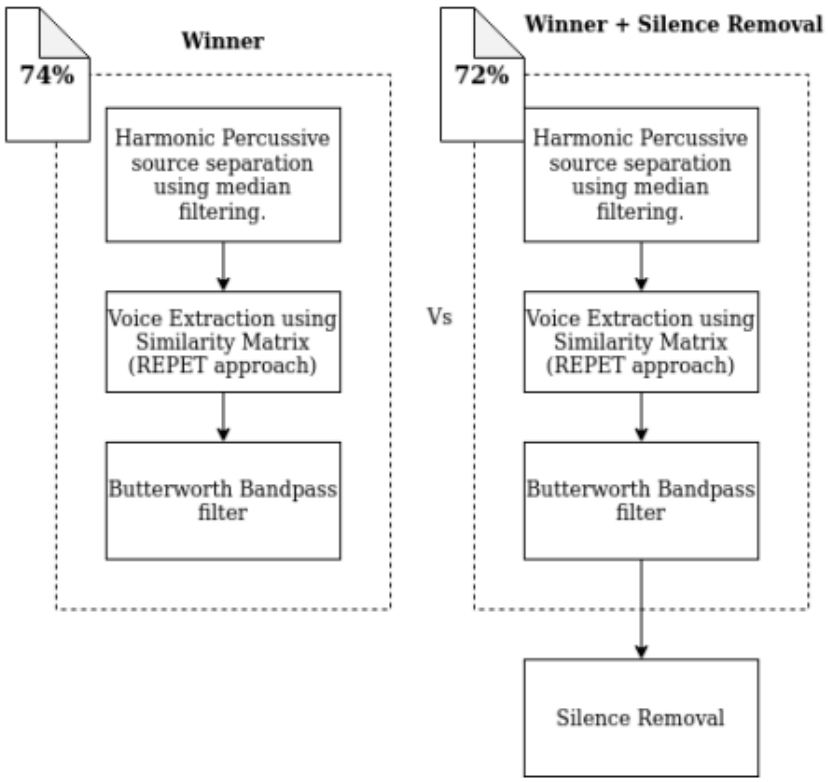

Fig. V. Evaluating the best voice isolation approach with winning approach against winning approach + Silence Removal

The silence removal had not been fruitful in voice isolation as it has decreased the performance of the winning approach. Why this has happened can be because even though the silence had been

removed from the start and the end, still there are remaining silence chunks in the middle of the songs due to the interludes.

The final signature generation approach had been established using the evaluation results. The Figure VI depicts the primary flow of phases in the digital signature generation.

The quantitative evaluation of this research project had been concluded with declaring the best approach to generate the digital signature by using REPET, Harmonic and percussive source separation and the band-pass filter.

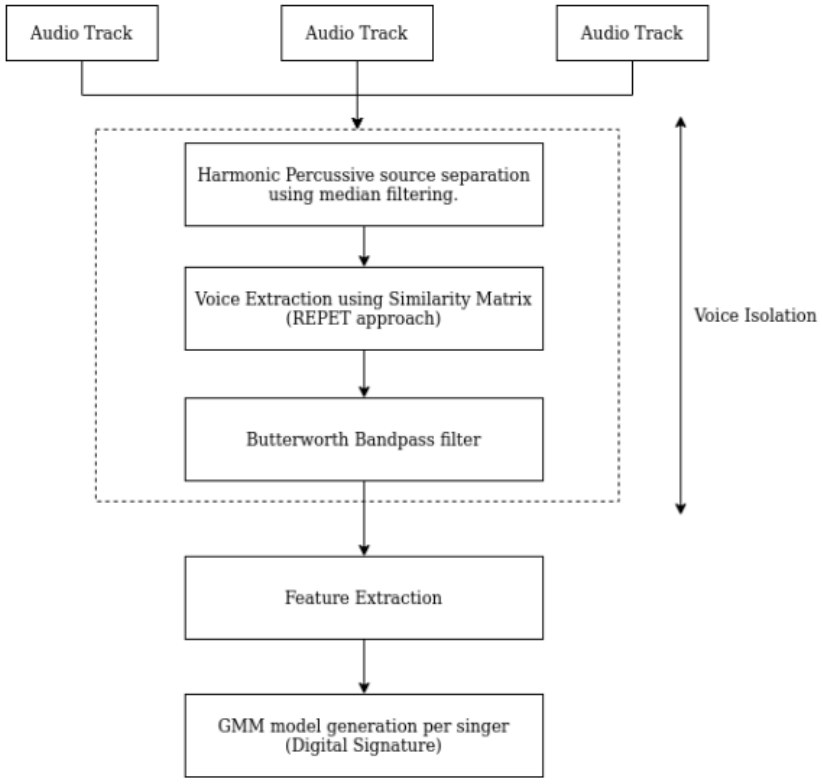

Fig. VI. Finalized research approach after evaluation

The qualitative evaluation, is performed using a pairwise evaluation. Several pairs of singers have been classified by considering the relationships given in the following list. Signatures for both the singers are generated and it is examined if the singer is identified correctly when given a song. The reason for pairwise evaluation is because this research focuses on reducing counterfeiting. In counterfeiting a song of a singer is sung by another person. So the signature

generated by the song will be compared with only the real singer's signature. So, a binary evaluation can be considered as the mostly suited evaluation approach for this project.

The pairs which had been made into consideration of this research are,

- Gender classification - Can the signatures of male and female singers be identified correctly?
- Classification of male singers - Can the signatures of different male singers identified properly?
- Classification of female singers - Can the signatures of different female singers be identified correctly?
- Classification of father son combinations - Can the signatures of a father and a son be identified properly?
- Classification of vocals of siblings - Can the signatures of sisters or brothers be identified properly?

4.4.1 Gender classification: An example for a pair of singers whose signatures were evaluated is Amal Perera with Deepika Priyadarshani. They had generated two different signatures and whenever a test song was provided, they were accurately identified as Amal Perera's or Deepika Priyadar- shani's. The following Figure VII shows how the results were generated.

```
In [6]: runfile('E:/4thYear1stSem/Final Year Research/Project/
Attempt1/featureExtraction/test_singer.py', wdir='E:/4thYear1stSem/
Final Year Research/Project/Attempt1/featureExtraction')
/amal/vocals10.wav
        detected as -  amal
/deepika/10.wav
        detected as -  deepika
```

Fig. VII. Artist signature identification

This evaluation had been done using other artists like, H. R. Jothipala, Nelu Adhikari, Jagath Wickramasinghe, and Anjaline Gunathilaka. All the voices were correctly identified as female and male when given a pairwise combination.

4.4.2 Classification of male singers: For this evaluation, singers whose voices are considered similar in Sri Lanka were chosen.

- Greshan Ananda and H.R. Jothipala
- Dayan Witharana and Jagath Wickramasinghe

Both the evaluations were successful. In both situations, test song was correctly classified. This observation has demonstrated that the signatures generated are robust and rigid even under similar features. Both combinations were correctly identified from each other.

4.4.3 Classification of female singers: This evaluation had also been done using female artist pairs who are said to have similar voices. The pairs which were considered in this project were,

- Neela Wickramasinghe and Nelu Adikari
- Anjaline Gunathilake and Latha Walpola

Both these evaluations had given positive results. All singers were correctly identified even though they sound similar.

4.4.4 Classification of father-son combinations: It is a fact that in Sri Lanka there are some father-son pairs who sound very similar. Therefore the proposed signature generation approach had been tested on their songs as well. The father- son pairs considered for the evaluation were,

- Milton and Ranil Mallawaarachchi
- Mervin and Amal Perera

The results of this evaluation were again accurately identified as the father or the son exhibiting the robustness of the signature. The signature generated shows the ability in extracting sensitive features from the vocalist.

4.4.5 Classification of siblings: The siblings vocals which were compared in this evaluation were,

- Umara and Umaria Sinhawansa

Both these singers have similar voices, and the signatures were unable to classify Umara from Umaria. Songs of Umara Sinhawansa were classified as Umaria Sinhawansa's while Umaria

Sinhawansa's voice was correctly identified as Umaria Sinhawansa's. The inability of this approach to identify these two singers cannot be answered directly but through some more experiments. It can be assumed that as these two are female singers and the voices are higher in frequency range, there might be issues when extracting features from them.

4.5 Further Evaluation

The following two criteria has been used to evaluate the signatures generated for the different singers.

- Two different signatures should be generated when the same song sung by two different singers.
- The same digital signature should be generated when different songs sung by the same singer.

The first condition had been tested by using the song" Ma sanasa" sung by both Mervin Perera and Amal Perera. They had specifically resulted in identifying the voices accurately as Mervin's and Amal's regardless of the test song being the same song. This experiment had given successful results.

The second condition had been tested by using two different songs of H.R. Jothipala. Both the songs were classified as H.R. Jothipala's songs discarding other signatures.

4.5.1 Discussion of qualitative approach: The pairwise evaluation had been successful in almost all the cases. It had shown just one negative result. The Table II below, summarizes the achievements of our research.

TABLE II

QUALITATIVE EVALUATION

#	Combination	Result
1	Gender classification	
2	Male singer classification	
3	Female singer classification	
4	Father son classification	
5	Sibling classification	X
6	Same song - different singers	
7	Same singer - different songs	

Evaluation of the research had been successful. It is safe to say that the generating signature is done using very specific features of voices.

5 CONCLUSION

Audio signal processing and music information retrieval has been used in this research to isolate vocals in songs, extract features from an audio signal and to generate unique models (signatures) for each artist. In this research, it was discovered that it is feasible to isolate voice from songs with music by using a combination of REPET with Band-pass Filter and Harmonic Percussive source separation. The signatures generated had been compared with other signatures to examine the sensitivity to capture specific features of vocalists. In all experiments, except one case, the singer had been identified correctly.

It was successful in storing the signatures as GMM models of size 22KB. For every singer this signature can be created using the proposed approach. The comparing process is easily done using the cpickle library of python.

This research is a novel research. The available literature treated voice isolation and singer identification as two different research problems. The voice isolation method proposed in our research is a combination of a usual voice isolation method REPET and audio analysis methods. It had shown a better accuracy than the usual REPET method. Signatures are represented as GMM models rather than HMM models. An evaluation of GMM model over HMM model is not discussed in this research but can be performed as future work.

This research has given a solution for identifying the imitated voices of artists and to locate counterfeit audio tracks. By using the proposed approach songs can be recognized either as sung by the original singers or by the counterfeiting singers. The qualitative evaluation proposed has demonstrated the ability to distinguish similar voices accurately. Consequently, the proposed technique can be used to find songs sung by the real singers.

In conclusion, this study has yielded a productive solution for identifying counterfeit songs with high accuracy.

6 FURTHER RESEARCH

The accuracy of the proposed approach could be increased further by experimenting with other signal processing techniques. The pre-processing steps used in this research used a combination of three signal processing techniques. This process could be modified and improved by experimenting with the other signal processing techniques. If voice isolation becomes more successful, it might improve the final result.

The Sri Lankan Sinhala songs have a close resemblance with Indian Hindi songs, where Sinhala songs with the exact melody of Hindi songs are common. Therefore, this research can be extended for Indian songs as well.

The representation model used as the signature in this research is a GMM model. This model can be replaced or refined further to obtain a better percentage accuracy than 74%.

REFERENCES

[1] J. O'Neal, "Advertising Agency's use of sound alike infringes singer's right of publicity," Available: https://www.theexpertinstitute.com/case- studies/advertising-agencys-use-sound-alike-infringes-singers-right- publicity.

[2] S. Froitzheim, "A Short Introduction to Audio Fingerprinting with a Focus on Shazam,", 2017.

[3] Velankar, Makarand, "Study paper for Timbre identification in sound," vol. 2, October 2003 Available: http://www.ijert.org/view.php?id=5942&title=study-

[4] K. Kristoffer Jensen, Timbre models of musical sounds - from the model of one sound to the model of one instrument, Technical report / University of Copenhagen / Datalogisk institut, 1999.

[5] A. Bonjyotsna and M. Bhuyan, "Signal processing for segmentation of vocal and non-vocal regions in songs: A review," ,February 2013, pp. 87-91.

[6] Velankar, Makarand. (2014). Automatic Classification of Instrumental Music Human Voice Using Formant Analysis. 2. 242. 10.13140/2.1.2926.6561.

[7] Sieranoja, Sami. (2019). How much k-means can be improved by using better initialization and repeats?. Pattern Recognition. 93. 10.1016/j.patcog.2019.04.014.

[8] Y. Li and D. Wang, "Separation of Singing Voice From Music Accompaniment for Monaural

Recordings," vol. 15, no. 4 pp. 1475- 1487,doi=10.1109/TASL.2006.889789, ISSN=1558-7924 May 2007 [IEEE Transactions on Audio, Speech, and Language Processing].

[9]Ozerov, Alexey and Philippe, Pierrick and Bimbot, Frederic and Gribon- val, Remi, "Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs," vol. 15, ,doi=10.1109/TASL.2007.899291, August 2007 [IEEE Transactions on Audio, Speech, and Language Processing].

[10]C. Hsu and D. Wang and J. R. Jang and K. Hu, "A Tandem Algorithm for Singing Pitch Extraction and Voice Separation From Music Accompaniment," vol. 20, no. 5 pp. 1482-1491,doi=10.1109/TASL.2011.2182510, ISSN=1558-7924,July 2012 [IEEE Transactions on Audio, Speech, and Language Processing].

[11] Zhu, Bilei and Li, Wei and Li, Ruijiang and Xue, Xiangyang, "MultiStage Non-Negative Matrix Factorization for Monaural Singing Voice Separation," vol. 21, pp. 2096-2107,doi = 10.1109/TASL.2013.2266773, October 2013 [Audio, Speech, and Language Processing, IEEE Transactions on].

[12] Rafii, Zafar and Pardo, Bryan, "A simple music/voice separation method based on the extraction of the repeating musical structure," vol. 21, pp.
221 - 224,doi = 10.1109/ICASSP.2011.5946380, June 2011 [Acoustics,
Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on].

[13] D. Dharini and A. Revathy, "Singer identification using clustering algorithm,", pp. 1927-1931,doi=10.1109/ICCSP.2014.6950180, April 2014 [2014 International Conference on Communication and Signal Processing].

[14] Deshmukh, Saurabh and G. Bhirud, "Analysis and application of audio features extraction and classification method to be used for North Indian Classical Music's singer identification problem," vol. 3, February
2014 [International Journal of Advanced Research in Computer and Communication Engineering].

[15] Y. Kim and B.Whitman, "Singer Identification in Popular Music Record- paper-for-timbre-identification-in-sound.
ings Using Voice Coding Features," vol. 3, September 2002

[16] Whitman, B. and Flake, Gary and Lawrence, Steve, "Artist detection in music with Minnowmatch," isbn = 0-7803-7196-8, doi = 10.1109/NNSP.2001.943160, February 2001.

[17] Fujihara, H. and Kitahara, T. and Goto, Masataka and Komatani, Kazunori and Ogata, Tetsuya and Okuno, Hiroshi, “F0 Estimation Method for Singing Voice in Polyphonic Audio Signal Based on Statistical Vocal Model and Viterbi Search,” vol. 5, doi = 10.1109/ICASSP.2006.1661260 June 2006.

[18] Meinard M"uller, “Harmonic Percussive Source Separation,”

[19] Sahoo, Tushar and Patra, Sabyasachi, “Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification,” vol. 6, doi = 10.5815/ijigsp.2014.06.04,May 2014 [International Journal of Image, Graphics and Signal Processing].

[20] Abubakar Sadiq, Abdulkadir and Othman, Nurmiza and Abdul Jamil, Muhammad Mahadi, “Fourth-Order Butterworth Active Bandpass Filter Design for Single-Sided Magnetic Particle Imaging Scanner,” vol.
10,June 2018 [Computers Electrical Engineering].

[21] Saini, Manish and Jain, Saurabh, “Comprehensive Analysis of Signal Processing Techniques Used For Speaker Identification,” ,May 2013.

[22] Giannakopoulos, Theodoros Pikrakis, Aggelos. (2014). Introduction to Audio Analysis: A MATLAB® Approach.

[23] Kos, Marko and Kacic, Zdravko Vlaj, Damjan. (2013). Acoustic classification and segmentation using modified spectral roll-off and variance-based features. Digital Signal Proc