The Nucleus ***52****, No. 1 (2015) 29-39*

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)

ISSN 2306-6539 (Online)

A Framework for Music-Speech Segregation using Music Fingerprinting and Acoustic Echo Cancellation Principle

F. Hussain*, H.A. Habib and M.J. Khan

Department of Computer Engineering, University of Engineering and Technology Taxila, Taxila, Pakistan

ARTICLE INFO

Article history :

Received : 19 January, 2015

Revised : 24 March, 2015

Accepted : 25 March, 2015

Keywords :

Source separation,
speech processing,
speech-music segregation,
spectral subtraction,
music filtering.

ABSTRACT

Background interference creates voice intelligibility issue for listener. This research work considers background music as interference for communication through smart phone in areas with loud background music. This paper proposes a novel framework for background music segregation from human speech using music fingerprinting and acoustic echo cancellation. Initially, background music is searched in the database by music fingerprinting. Identified background music is registered and segregated using acoustic echo cancellation. Proposed approach generates better quality music speech segregation than existing algorithms. The research work is novel and segregates background music completely in comparison to existing approaches where single instruments are segregated successfully.

1. Introduction

Interferences can occur for multiple reasons in the environment. This research work considers the scenario of two people communicating through mobile or smart phone. Background music is considered as interference. This causes problems in hearing for the receiver and poses speech intelligibility issue. This research work has the objective to remove this background music at the sender side so that only human voice is sent to receiver.

Music is a complex signal and it has various representations: pitch for melody and harmony; rhythm for tempo, meter and articulation; dynamics for timbre and structure. Music signal possesses specific acoustic and structural characteristics that does not exist in spoken language or environment noises. Human auditory system has capability to segregate speech from background interferences. It is challenging for an intelligent system to segregate speech from background music.

Music speech segregation algorithms have objective to segregate human speech from background music. Researchers put more focus on separating musical instruments, source separations and less focus on separating human speech from music. These algorithms can be classified according to type of segregation they are performing. Some algorithms report segregation human speech and music. Other algorithms report segregation of instruments and filtering particular instrument voice in output stream. Some algorithms target further e.g. separating tones with different timbers [1].

2. Related Work

Music speech segregation algorithms are based on computational auditory scene analysis, blind source separation. Algorithms targeting segregation of human speech considers behavior of human speech. Human Speech mainly consists of two parts, voiced and unvoiced speech. Researchers reported that 20 to 25% of the human speech is unvoiced [2]. Voiced speech is harmonic in structure while unvoiced speech is inharmonic in structure. Voiced speech is comparatively easy to segregate it from background compared to unvoiced speech segregation. Various model based approaches like Gaussian mixture model (GMM), Hidden Markov Model (HMM), and Nonnegative matrix factorization (NMF) [3-5] were used for separation purpose, but these approaches mainly base on speaker information. Model based approaches requires whole speech model including voiced and unvoiced portion along with the speaker identity sometimes and speaker model quite often. In order to separate unvoiced speech in speaker independent way from the voiced type, an algorithm was presented in [6].

Different methods are proposed in literature for unvoiced speech separation. For instance in computational auditory scene analysis (CASA), methods based mainly on feature extraction are proposed. For segmentation purpose of speech, onsets and offsets are used [2], however they don't help in differentiating voiced and unvoiced speech. One of the important algorithm, for unvoiced speech separation used in CASA, named as spectral subtraction [7]. For grouping of voiced speech, 6-dimensional pitch based features (6F) were used in algorithm presented in [8].

* Corresponding author : fawad.hussain@uettaxila.edu.pk

Ideal binary mask (IBM) can be used as classifier to formulate speech separation. An algorithm based on supervised learning technique using 6F [10] or amplitude modulation spectrum (AMS) features was proved to be impressive in single speech segregation. Hybrid approaches were also used recently [9], which trains support vector machine (SVM) for classification using 6F and AMS features, it showed considerable improvement in performance. Different features like pitch-based, gammatone frequency cepstral coefficients (GFCC) and AMS features are used for classification purpose of T-F units in [6].

Compared to singing voice separation, speech separation is more complex with more pitch variations. Moreover speech separation from background accompaniments is more complex as compared to the singing voice separation from the background because in separating the source of our interest from different types of background noise, most probably having varied characteristics like whether it can broadband or narrowband, periodic or aperiodic. More complexity is added to the problem by the fact that speech and background noise are independent of each other mostly with uncorrelated spectral contents.

There are some existing methods that were used for monaural speech separation from background noise, they are usually categorized, into three types based on their working principles, as: model based methods [11-12], spectrogram factorization [12-13] and pitch based methods [14-15]. There are few limitations with these methods as well. Spectrogram factorization encounter problems in assigning components to relevant sound sources. As the number of musical instruments increases, its performance decrease significantly. One of the drawback with model based approaches is that they require fair amount of only music chunks in order to model the characteristics of background music. Pitch based methods have fewer limitations as compared to the previous two. Speech signal's pitch contours are its only requirement that can easily be extracted from a mixture of very short duration and it does not need any only music accompaniment parts.

Monaural speech segregation is comparatively a difficult task due to the availability of the single channel recording only, the challenge that one face is that it do not provides source spatial information which is normally in binaural situations. In case of monaural speech segregation, only intrinsic properties of speech are there to work with, intrinsic properties of speech are its harmonic structure and onsets [16]. Research using the said features has obtained quite advancement in voiced speech segregation in vulnerable situations [17-20]. On the other hand if we talk about unvoiced speech segregation, this problem is still a challenge.

Speech enhancement is proposed to enhance the distracted/noisy monaural speech [21]. Different algorithms that are used for enhancement of speech includes Wiener filtering, minimum mean square error based estimator, spectral subtraction and subspace analysis. These methods are quite good in dealing with unvoiced speech. The drawback with speech enhancement algorithms is their presumptions about statistical properties of noise that distract speech; it reduces the ability of these algorithms to deal with general interference, e.g. they often assumes that the noise which is distracting the speech is stationary which is clearly not true if we talk about the real world scenarios where interference often changes abruptly.

Model-based separation systems represents another class of techniques, their main emphasis is on source pattern modeling and they reduce the separation to an estimation problem. In parallel to this a form Gaussian mixture is used as a composite source model [22]. Unvoiced speech can be segregated significantly using model-based approaches but there assumptions about mixture only having speech utterances reduce their ability.

The central idea of computational auditory scene analysis (CASA) is to obtain sound organization based on perceptual principles [23]. CASA comprises of two stages; segmentation and grouping. The input is broke down into segments during segmentation phase, each segment is originates from the single sound source. After segmentation the next phase is grouping, which mainly group the segments from segmentation process that comes from the same source into a stream. Ideal binary masking (IBM) is suggested as CASA's main goal [24]. Subject tests were carried out to prove the performance of IBM, it proved that IBM leads to great improvement in the performance of hearing from both normal as well as hearing-impaired people [25-27]. Noise type is incorporated with spectral subtract for unvoiced speech segregation [28]. In [6] the idea of segmentation based on spectral subtraction is presented.

Segregating speech from such distractions is very useful in different applications. Earlier in the research, more emphasis was on voiced speech with less attention on unvoiced speech segregation. In [29] monaural unvoiced speech segregation from non-speech interference was studied. Main motivation for it was the Bregman [16] theory of auditory scene analysis. Unvoiced IBM is directly estimated in an algorithm presented in [29]. Initially the voiced binary mask is estimated using supervised learning approach [30]. It requires two stage process, after the first step. While performing unvoiced speech segregation.

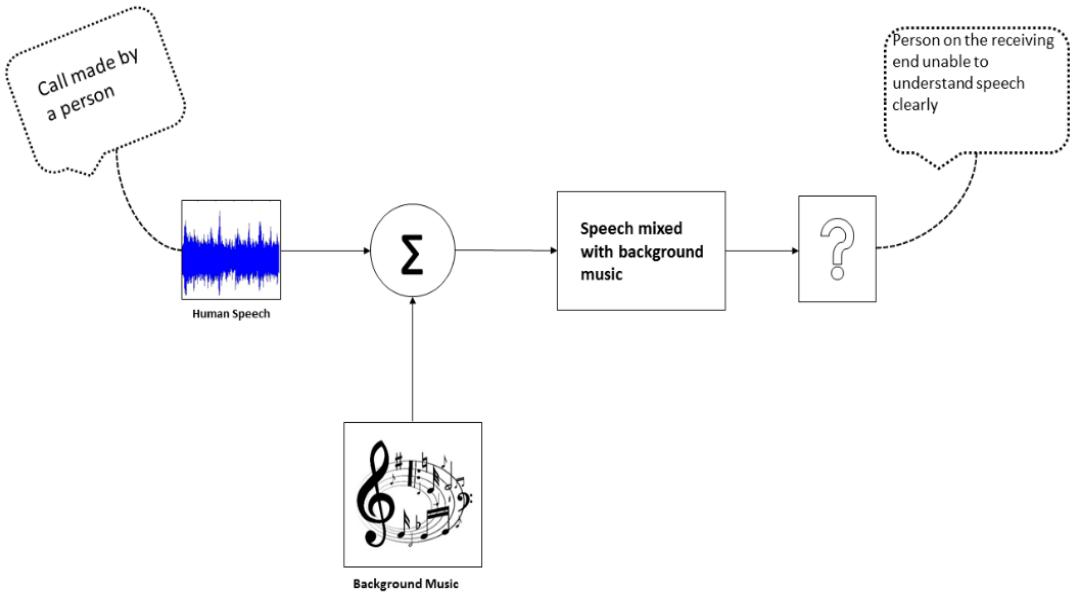

Fig. 1: Problem formulation

During the first stage i.e. segmentation the noise energy estimation is done using voiced binary mask and afterwards spectral subtraction is done to generate the T-F segments of unvoiced speech.

It is also important to note that noise has different characteristics in different environments, different grouping techniques were proposed depending upon the noise type [29]. Based on the variation in the noise energy, it was categorized into three classes: stationary, non-stationary and highly non-stationary. For each of these noise types, a relevant grouping technique was applied for grouping the target segments.

The rest of the paper is organized in the following sequence; section 2 describes the methodology, section 3 explains the experimental setup, section 4 shows the results followed by conclusion and future work.

3. Methodology

3.1 Problem Formulation

Fig. 1 shows the problem formulation model. The basic theme of this research revolves around the mobile phone communication. Real time scenario is, a person call to some other person on his mobile phone and let's suppose that there is a music in the background environment of a calling person so when the called person receives a call, he will face difficulty in understanding the calling persons speech due to the interference from his musical background. The interference in the background may be due to anything like moving car, hammering etc. but for instance we are considering musical background as an interference/ noise. Mathematically it can be given as;

$$C_S = H_S + B_m \quad (1)$$

Where C_s is the calling person who when calls to some person, the information that is transmitted to the called person through that call is sum of human speech (H_s) and background music (B_m) which is designated as interference or noise in this case.

3.2. Proposed Solution

A novel approach is proposed in this research which helps to remove the background music quite efficiently and conveying the human speech cleanly. When a person calls to some other person on his mobile phone and consider there is a noisy background (in this case music is considered as noise only) for a calling person. The musical background makes the communication difficult between two parties. The solution to this problem is proposed in this research. There are lot of intelligent music applications available from smart phones, like SHAZAM, SoundHound, TrackID, mobion music global, musiXmatch Lyrics Player and Echoprint. Echoprint is used in this research as music identification application. The reason for using Echoprint is that this application is as good in performance rather better in some cases as compared to applications mentioned above, more importantly Echoprint is an open source [30] application and anyone can use it to build music fingerprinting into their system. Echoprint is quite fast, it identifies the song in less than one tenth of a second and also there are currently more than one million songs in the database that can be identified by it and the song database is increasing day by day.

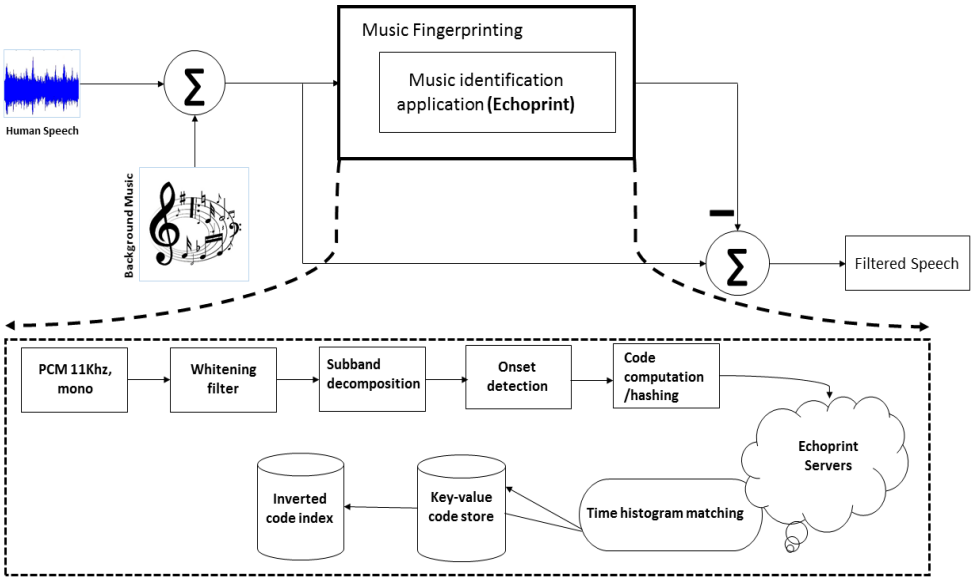

Fig. 2: Proposed system model

Whenever a person calls to someone on his phone and there is a background music in the environment of calling person, upon receiving the call by called person, the speech from the calling person is mixed with the background music and it will make difficult for a called person to understand the speech, the music identification application (Echoprint) is integrated with our system and the initial chunk of the audio mixture is fed to the application which immediately recognize the music and it is downloaded on the mobile phone and this music will be subtracted chunk by chunk from the audio mixture and as a result the system will be left with only human speech that the calling party is listening. This proposed algorithms system diagram is shown in the Fig. 2.

4. Experimental Setup

4.1 Algorithmic Details

The flow chart of the problem at hand is shown in Fig. 3. As it is mentioned that the algorithm is proposed with respect to the specific scenario i.e. mobile phone communication, it is clear from the flow chart shown below In Fig. 3 that whenever a call is originated from the transmitter (here the calling party is called as transmitter) to the receiver (the called party) and their might be some music in the background of the transmitter so in that situation it will be difficult for a receiver to understand that speech of a transmitter due to the interfering background music.

The flow chart of the proposed algorithm which is the solution for the above mentioned problem is shown below in Fig. 4.

For experimentation purpose we have created our own dataset by using speech and music audio provide by GTZAN music-speech dataset. For creating the dataset different types of music is used like western, eastern, Arabic, and news channels background music etc. 30 different people that include males and females who have volunteered to help for recording purposes as to add human voice with the background music. 200 audio clips (30 seconds duration) have been created in the dataset for the testing purposes of the proposed algorithm.

It is important to mention here that both the audio streams that are being manipulated must have the same sampling rate, after equalizing the sampling rate of audio stream further there are two scenarios while providing the solution for the problem at hand.

4.1.1 Ideal Case

First scenario describes the ideal case in which both the audio streams which are being operated upon, i.e. the background music audio and the other one is the mixture of background music and human speech, are synchronized in terms of start time or in other words we can say that both the audios are aligned sample-to-samples. Fig. 5 shows the above mentioned scenario in which both the audio streams are aligned to each other. In this case the background music is separated from the human voice completely by simply subtracting the background music audio from the mixture and as a result only human voice is transmitted to the receiver side despite of the interfering background music. The function is given by;

$$O_a = M_{m+v} - B_m \quad (2)$$

Where O_a is output audio, M_{m+v} is mixture of music and human voice and B_m is background music audio.

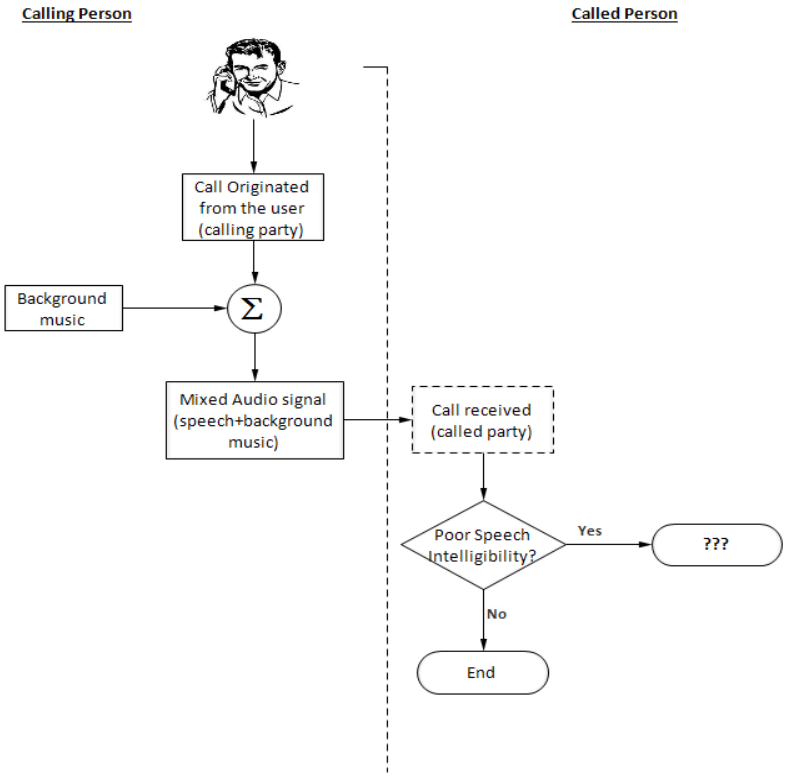

Fig. 3: Flow chart of a problem.

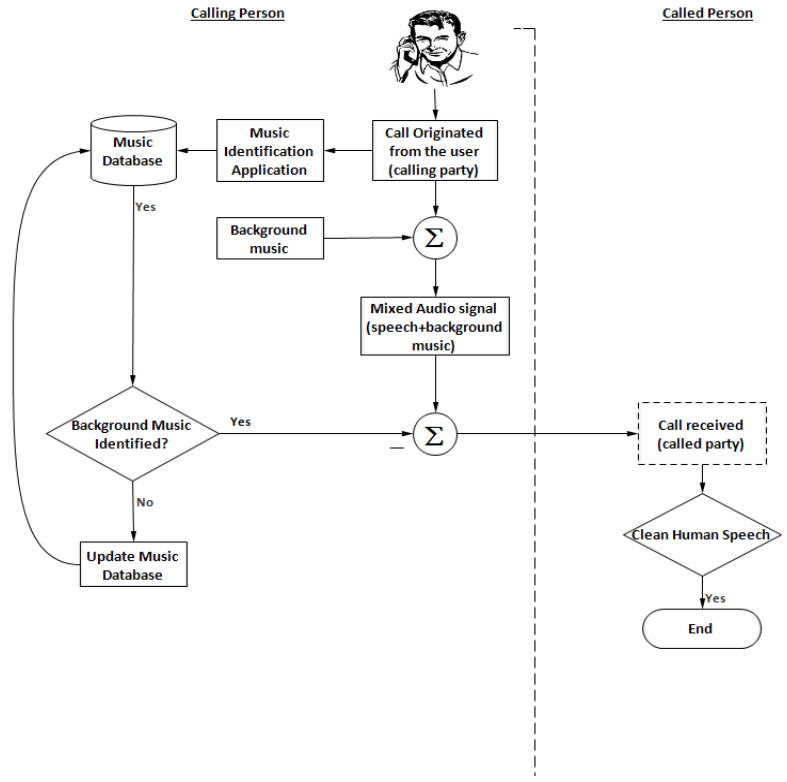

Fig. 4: Proposed algorithm flowchart

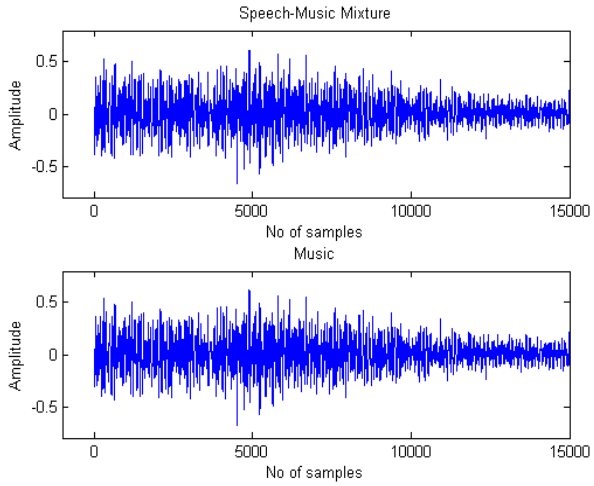

Fig. 5: Ideal case (both audio streams are aligned sample wise)

4.1.2 Sample Misalignment

This scenario is somewhat tricky and complex, as compared to the first one, in which both the audio streams are not aligned completely sample wise so it's impossible to perform the subtraction operation between the audio streams as mentioned above in first scenario. This problem is shown below in Fig. 6.

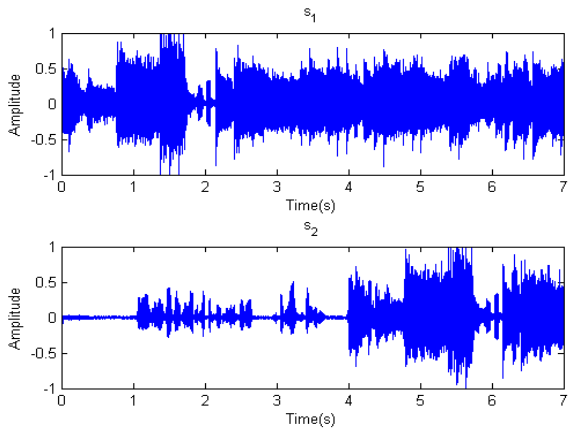

Fig. 6: Sample-to-sample misalignment between two audio streams

This problem is addressed by time delay calculation between two signals, i.e. the background music signal and the signal in which music is mixed with human speech, by cross correlation between two signals which gives us the background music signal leading or lagging in time as compared to the mixed signal. Fig. 7 shows the process of correlation, as we can see that in the upper half of the figure the signal is lagging compared to the reference signal and in the lower half the signal is leading. The peaks in the following figure shows that the signal that we are trying to find is present and starting from the sample where the peaks are maximum.

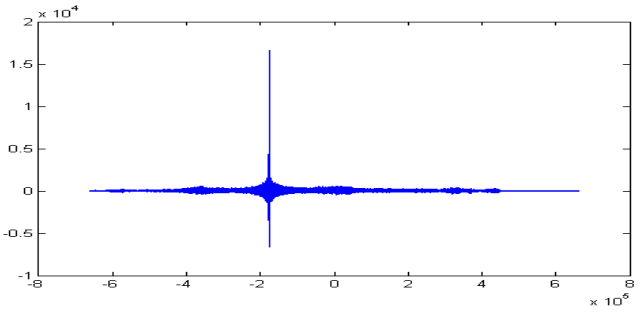

Fig. 7: Correlation between signals (finding lead/lag time of signal w.r.t reference signal)

After the cross correlation between the signal, the signals are aligned in time sample-by-sample as shown in Fig. 8. Now the subtraction can be performed between two signals and as a result only human voice is left which is being transmitted to the receiving end side.

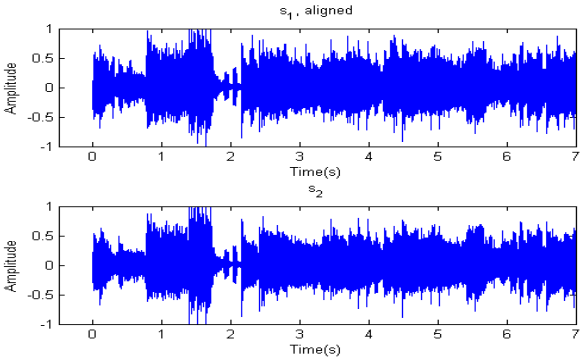

Fig. 8: Sample-to-sample aligned signals after cross-correlation

4.2 Music Identification System

Music identification application is a very important part of the proposed algorithm, which helps in finding the background music. As shown in the Fig. 9 whenever a person A calls to a person B, for instance consider there is a background music in the environment of person A, the proposed solution is available on person A's mobile phone, music identification application identifies the background music within 10^{th} of a second and is downloaded on the phone which is subtracted from the mix signal (human voice and background music) and as a result only clean human voice is transmitted to the person B's phone.

The music identification application which is being used here is available off the shelf and completely open source named as Echoprint [16]. The lower portion in the figure inside the dashed line rectangle shows the architecture of Echoprint. Echoprint listens to audio from any source like phone, computer or environment and finds out what song it is. It works very fast and with a very good accuracy. It is so robust that I can identify the noisy versions of the original or rerecorded version using some low quality recorders with strong outside interfering sources.

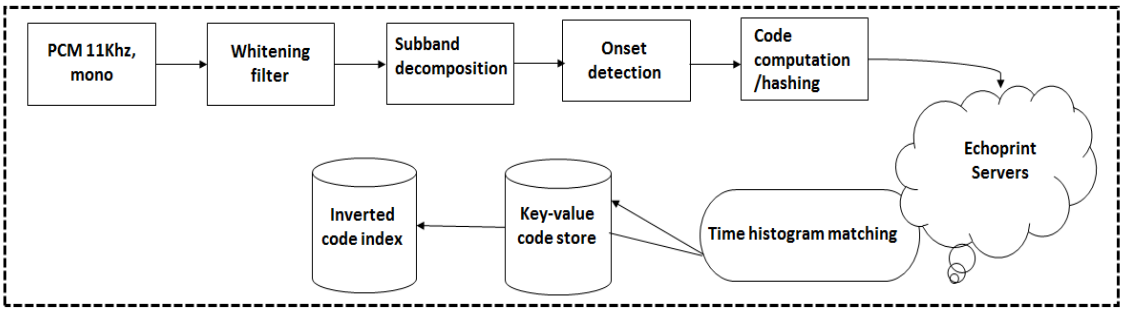

Fig. 9: Echoprint (music identification application) [30]

4.2.1 Fingerprinting

Audio fingerprint [31, 32, and 34] is a condensed digital summary produced from an audio signal which can be used to identify/locate the music sample from a database. It is used quite frequently since last decade or so and became an important feature for a large scale music identification services. Echo Nest Musical Fingerprint (ENMFP) [31] was presented earlier, it analyze music audio in detail using Analyze service of Echo Nest [34],and mostly the successive segments of fingerprints were based on chroma vectors. Its effectiveness was proved by identifying the same track with different encodings, but was unable to handle severe spectral distortions like over the air recordings. Another drawback of it is its dependency on Analyze process output and if the Analyze process had not been carried out already, it becomes computationally very expensive. Later another music identification service, named as Echoprint, was presented which overcomes most of the drawback of ENMFP. Chroma features were not used in it, it only relied on the timing of successive beat-like events, this feature is quite important and the best thing about it is that it is quite robust under wide channel range and noise. Another advantage of Echoprint, in order to perform onset detection, it has its own scheme which is quite simple as compared to the one in Analyze.

4.2.2 Code Generation from Audio

In order to achieve robustness in the presence of over the air (OTA) recordings and spectral changes, Echoprint depends only on relative timings between successive beat-like onsets detected in audio. Before the further analysis the signal is whitened in order to improve the robustness against variable over-the-air channel characteristics. Whitening is achieved by applying inverse (FIR) to the signal but before doing this an estimation of 40-pole LPC filter is done by taking the autocorrelation of the 1 sec block of the signal. In this way any sort of stationary resonance in the signal that may arise from speaker, microphone or room in OTA will be reduced [35].

After the application of whitening filter to smooth the signal, the next step is a subband decomposition into 8 bands. 8 band subband decomposition is done for the search purpose of onsets. 1onset/second per band is the Echoprint's target onset rate. Pairs of IOI (Inter-onset-intervals) in each band are combined to make a hash. In order to make the onset detection robust and against missed or false onsets, four successors are also considered along with each onset. By considering all possible successive onsets from the four, six different hashes (IOI pairs) are created. So if we would like to find the overall hash rate, it will approximately be like: 8 (bands) x 1 (onset/sec) x 6 (hashes per onset) ≈ 48 hashes/sec. As onsets are approximately 1 second a part so the order of quantized IOIs is 1/0.0232 = 43 or 5-6 bits, and a pair of onsets comprised of 12 bits of information. After that 3 bits of band index are combined with it to generate raw hash that is stored with its occurrence time within the file [35].

4.2.3 Code Matching to find Songs

There is a database of more than 1 million music tracks as its still in its early age but its growing rapidly both by the users and administrators. Matching activity works in a simple way, the input in it is an unknown query Q, based on the input Q, the required track is to be found in the database. In order to build the database each track is decomposed into 60 second segments and adjacent sections overlapping by 30 seconds. In this way the bias is removed which comes when longer songs provide more matches for the set of query hashes. Document D is the term normally used for the codes of 60 seconds segment in an inverted index. Document ID is the combination of unique track ID and the segment number. Echoprint uses the Apache Solr server as a data store. When the query is made, the server returns the documents with maximum number of matches for each code term in the query. It is important to note that there is rarely only one document which contains the maximum code matches than other documents indexed. Echoprint use top 15 matches to find out the corresponding track as it is believed that if the track exists in the database, it will be in those top 15 matches. Later time histogram matching is done and they use top two histogram buckets to inform the "true score". This ensures that the code occurs in correct order no matter even if Q is from the different section of the song [35].

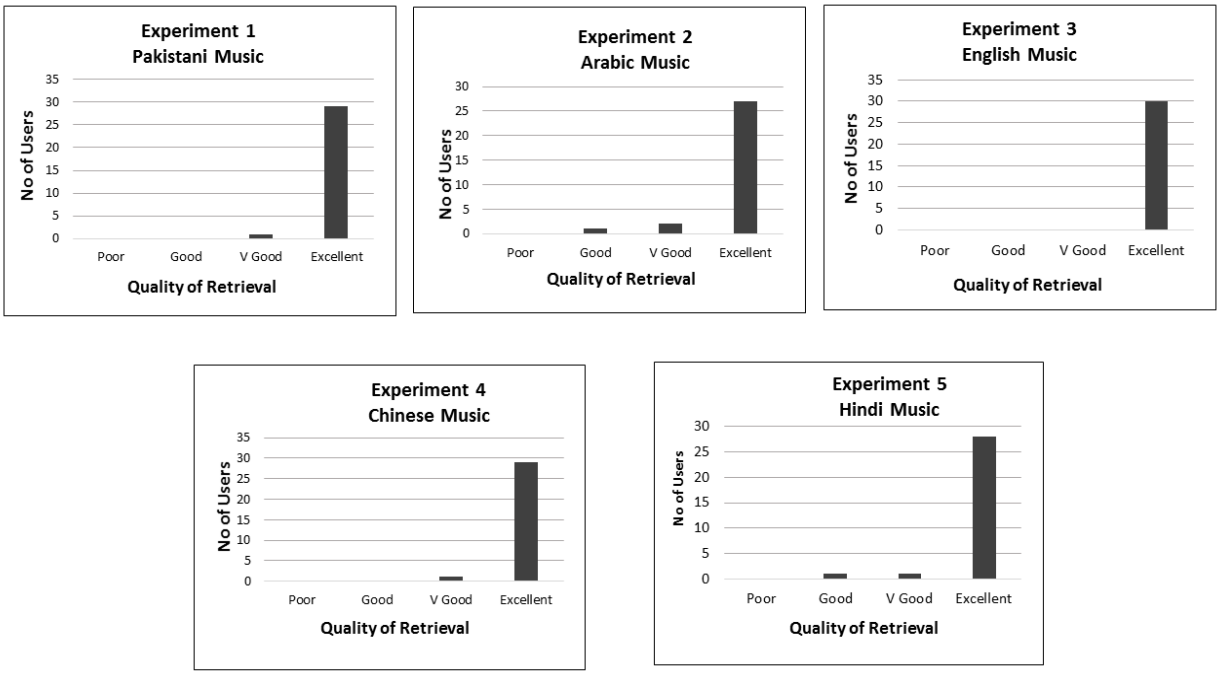

Fig. 10: Experimental results

All possible documents are arranged on the basis of their true score. If multiple documents from the same track exists in the list, all the documents are removed except the one with the highest score. The document present at the top of the list, with its true score significantly higher than all other documents in the list, returned as positive match. On the other hand if the gap between true scores of the top two results in the list is insignificant, it means there is no match.

5. Experimental Results

200 audio files, of different types of music, are over all used for the subjective evaluation of the proposed algorithm with 40 audio files of each type of music (Pakistani, Arabic, English, Chinese and Hindi), with the duration of 15 seconds each. 30 Males and females speakers are randomly chosen for the recording of the complete dataset. 30 different males and females volunteered for the evaluation purpose as a result of which the quality of the proposed algorithm is deduced. Results came out from subjective evaluation by performing multiple experiments are shown in Fig. 10.

Table 1 shows the cumulative result of subjective evaluation of all the individually performed experiments, it is also evident from Fig. 11 that the proposed framework achieves a high percentage (95.33%) of accuracy of speech separation from background music.

In addition to the subjective analysis of our results as shown above, the results of the proposed algorithm are also

Table 1: Overall evaluation result of all experiments

Experiments	Poor	Good	Very Good	Excellent
1	0	0	1	29
2	0	1	2	27
3	0	0	0	30
4	0	1	1	29
5	0	1	1	28

Computed with respect to the volume variations of the background music. It is noted that in the real time environment mostly the background music that is interfering with the person's speech talking on the phone is at some distance so as a result the strength of the music signal which is being added and transmitted toward the listener on the other side of the phone is somewhat weak as compared to the music signal if it is in the close vicinity of the person talking on phone or in simple words it can be said as the music signal has a low volume.

Results, shown in the Fig. 12 regarding volume variation of the background music, shows that the human speech is well understood on the receiver side (called party). We can see in the third subplot of the above figure that considerable amount of music is removed from the mixed signal even when the volume of the background music is lowered by 20%.

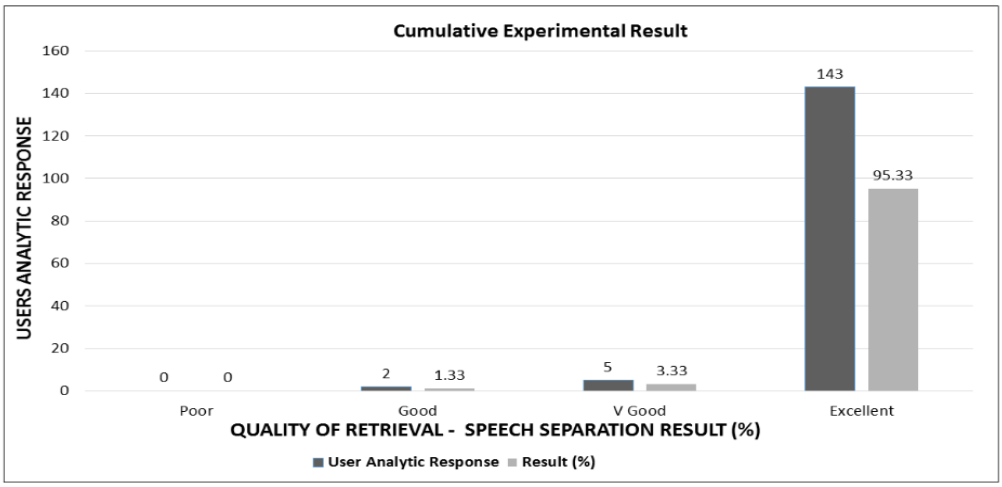

Fig 11: Cumulative result

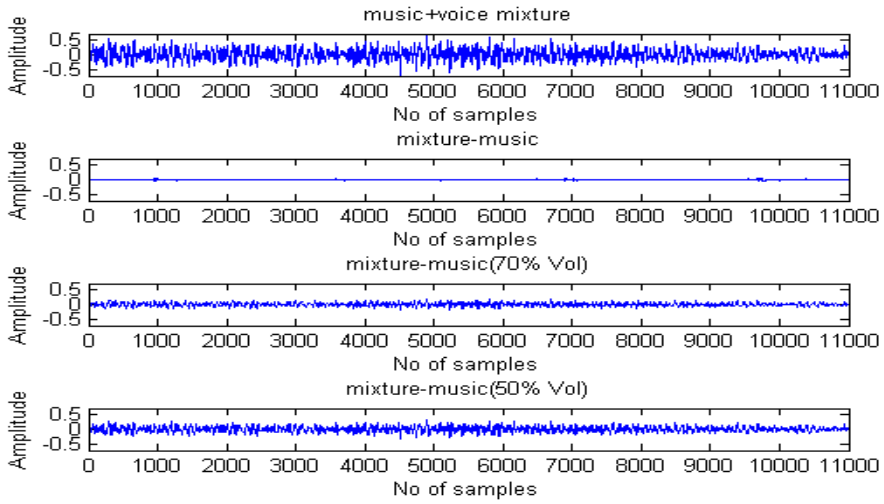

Fig. 12: Volume effects on speech music segregation

In order to reduce this problem even more and to make the proposed algorithm more robust, another solution is provided as well and it gives better results at the cost of a very small delay. As explained in the experimental setup that whenever a person calls to some other person and if there is a music in the background of calling person, it interferes with the speech and the called party cannot hear the speech clearly. This algorithm use the music identification application (Echoprint) in this case to immediately (less than 10^{th} of a second) identifies background music and download it to the device, the small delay that was mentioned above is here and it is before subtracting the downloaded music signal from the mixed (music speech) signal, the volume of the downloaded signal is reduced by 20% and then the signal is subtracted from the mixed signal and in this way the background music is separated from speech and suppressed up to more than 90% and the called party can clearly understand the speech. Results are shown below in Fig. 13 and in comparison to the Fig. 12 shown above are more promising and robust.

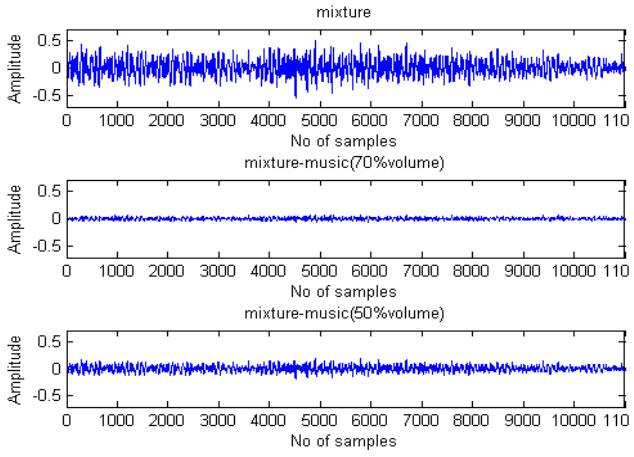

Fig. 13: Reduced Volume effects on speech music segregation

6. Conclusion and Future work

The algorithm proposed in this research shows quite promising results for speech-music segregation problem. This proposed algorithm is quite fast and robust as compared to the previously implemented systems in the literature. One of the important reason for the algorithm to be fast is that it works in the time domain as compared to the frequency domain where delays comes due to the intermediate filterbank implementations and manipulations.

Previously these kind of systems were implemented using hardware technologies i.e. DSP processors or FPGA chips to be used with actual embedded systems. The proposed framework is state of the art in itself because it utilizes the available computational equipment within smart phone instead of using extra hardware and provides with excellent results.

There are still some points in this algorithm on which further work can be carried out in future, for example the assumption that we took in this algorithm is that the background music always starts from its zero time while someone conversing on a phone. The extended work that will be carried out in future is to assume that the background music can be at any point, mean the song that is running in background can be at any point between its start and end time so in that case the first task to be done is to be identify the current instant of time for that music signal and then the subtraction maybe carried out from that point. The next point that can be worked out in future is the volume variation of the music signal and a mechanism can be devised to control the volume of the downloaded music signal by the music identification application in such a way (may be by means of averaging, by finding out normally how reduced the volume can be for the real time background music) so that the music can be separated entirely from the speech.

Acknowledgment

This work has been supported fully by the Directorate of Advanced Study and Research (ASR&TD) University of Engineering & Technology Taxila. I would like to thank my friends and colleagues who has helped me in any capacity in this research work.

References

[1] Y. Fukayama, D. Tanaka and T. Kataoka, “Separation of individual instrument sounds in monaural music signals by applying statistical least-squares criterion”, International journal of Innovative Computing, Information and Control (IJICIC), vol. **8**, March 2275-2283, 2012.

[2] G. Hu and D.L. Wang, “Segregation of unvoiced speech from non-speech interference,” J. Acoust. Soc. Am., vol. 124, pp. 1306–1319, 2008.

[3] J.R. Hershey, S.J. Rennie, P.A. Olsen and T.T. Kristjansson, “Super-human multi-talker speech recognition: a graphical model approach”, Comput. Speech Lang., vol. 24, pp. 45–66, 2010.

[4] A. Reddy and B. Raj, “Soft mask methods for single-channel speaker separation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 6, pp. 1766–1776, 2007.

[5] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 1, pp. 1–12, 2007.

[6] Hu, Ke, and D. Wang. "SVM-based separation of unvoiced-voiced speech in cochannel conditions." Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, Kyoto, Japan, pp. 4545-4548, 2012.

[7] K. Hu and D.L. Wang, Unvoiced speech segregation from non-speech interference via CASA and spectral subtraction”, IEEE Trans. Audio, Speech, Lang. Process., vol. 19, pp. 1600– 1609, 2011.

[8] G. Hu and D.L. Wang, “A tandem algorithm for pitch estimation and voiced speech segregation,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, pp. 2067–2079, 2010.

[9] Z. Jin and D. L. Wang, “A supervised learning approach to monaural segregation of reverberant speech,” IEEE Trans. Audio, Speech, Lang. Process., vol. 17, pp. 625–638, 2009.

[10] K. Han and D. L. Wang, “An SVM based classification approach to speech separation”, ICASSP, 2011, pp. 4632–4635.

[11] Y. Li and D.L. Wang, “Separation of singing voice from music accompaniment for monaural recordings”, IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4, pp. 1475–1487, May 2007.

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

[13] B. Raj, P. Smaragdis, M. V. Shashanka, and R. Singh, “Separating a foreground singer from background music”, Proc. Int. Symp. Frontiers Res. Speech Music (FRSM), Mysore, India, 2007.

[14] S. Vembu and S. Baumann, “Separation of vocals from polyphonic audio recordings”, Proc. ISMIR, pp. 337–344, 2005.

[15] T. Virtanen, A. Mesaros, and M. Ryynänen, “Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music”, Proc. SAPA, Brisbane, Australia, 2008, pp. 17-22. 2008.

[16] A. Bregman, “Auditory Scene Analysis”, Cambridge, MA: MIT Press, 1990.

[17] G.J. Brown and M. Cooke, “Computational auditory scene analysis”, Comput. Speech Lang., vol. 8, pp. 297–336, 1994.

[18] G. Hu and D.L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” IEEE Trans. Neural Network, vol. 15, no. 5, pp. 1135–1150, Sept., 2004.

[19] Z. Jin and D.L. Wang, “A supervised learning approach to monaural segregation of reverberant speech,” IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 4, pp. 625–638, May 2009.

[20] P.Li, Y. Guan, B.Xu, and W. Liu, “Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 6, pp. 2014–2023, Nov. 2006.

[21] Loizou, Philipos C. “Speech enhancement: theory and practice”. CRC press, 2013. ISBN 9781466504219

[22] M.H. Radfar and R.M. Dansereau, “Single-channel speech separation using soft masking filtering,” IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 8, pp. 2299–2310, Nov. 2007.

[23] D. L.Wang and G. J. Brown, “Computational Auditory Scene Analysis: Principles, Algorithms and Applications”, Eds. Hoboken, NJ: Wiley-IEEE Press, 2006.

[24] D.L. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” Speech Separation by Humans and Machines, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.

[25] D.S. Brungart, P.S. Chang, B.D. Simpson and D.L. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time–frequency segregation," J. Acoust. Soc. Amer., vol. 120, pp. 4007–4018, 2006.

[26] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary- masked speech: Implications for noise reduction," J. Acoust. Soc. Amer., vol. 123, pp. 1673–1682, 2008.

[27] D.L. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time–frequency masking," J. Acoust. Soc. Amer., vol. 125, pp. 2336–2347, 2009.

[28] K. Hu and D. L. Wang, "Incorporating spectral subtraction and noise type for unvoiced speech segregation", Proc. IEEE ICASSP, 2009, pp. 4425–4428.

[29] Hu, Ke, and D. Wang. "Incorporating spectral subtraction and noise type for unvoiced speech segregation." Acoustics, Speech and Signal Processing, (ICASSP), pp. 4425-4428, April 2009.

[30] G. Hu, "Monaural speech organization and segregation", Ph.D. dissertation, Ph.D. dissertation, Biophysics Program, The Ohio State University, 2006.

[31] http://echoprint.me/

[32] Daniel P.W. Ellis, B. Whitman, T. Jehan, and P. Lamere, "The Echo Nest musical fingerprint", Proceedings of the International Symposium on Music Information Retrieval, Aug, 2010.

[33] J. Haitsma and T. Kalker. A highly robust audio fingerprinting system with an efficient search strategy. Journal of New Music Research, 32, no. 2, pp. 211–221, 2003.

[34] T. Jehan, "Creating music by listening", PhD thesis, Massachusetts Institute of Technology, 2005.

[35] A. Wang, "An industrial strength audio search algorithm" International Conference on Music Information Retrieval (ISMIR) Baltimore, Oct. 26–30, 2003.

[36] Ellis, Daniel PW, B. Whitman, and A. Porter, "Echoprint: An open music identification service." ISMIR 2011 Miami: 12th International Society for Music Information Retrieval Conference, October 24-28. International Society for Music Information Retrieval, 2011.