Asian Journal of Information Technology 15 (23): 4770-4782, 2016
ISSN: 1682-3915

Multichannel Speech Processing and Separation Using Hybridized K-Superset Heuristic Algorithm

[1]Logeshwari and [2]Anandha Mala
[1]Department of Information and Technology, St. Joseph's College of Engineering, Chennai, India
[2]Department of Computer Science and Engineering, Eswari Engineering College, Chennai, India

Abstract: The presence of competing speakers signal input mixture in a noisy and fluctuated environment greatly degrades the performance. Noise and fluctuations hinders the process of speech separations. Very few research works accounts for these factors in their supervisory and unsupervisory estimation methods. The limitation of libraries or known speech samples and its recognition in a new environment makes speech signal processing a cumbersome exercise. An intelligent unsupervisory method would be a better choice for such requirements. Keeping view of this challenge, this research study proposes a novel method to separate multichannel speech signal from a single mixture captured in both stationary and non-stationary noisy and fluctuating environment using Multichannel B Hybridized K-superset Heuristic Speech separation Algorithm (MC-HKHSA). MC-HKHSA estimates pitch values for voiced and fluctuated voiced segments and forms supersets for multiple speakers. Noisy segments are filtered and fluctuated voice segments are grouped as a unique stream. This approach involves coarse and fine level speech processing and segregation mechanism making the overall process hybridized. Aim of MC-HKHSA is to segregate individual speech signals retaining its intelligibility, quality and naturalness. It removes excess background residual noise while retaining the positive features of enhanced speech. Simultaneous process management reduces the error rate due to inter-algorithmic value conversions. Simulation and experimental evaluations demonstrate that our approach outperforms other existing schemes with various energy levels. The improvement was due to our fine algorithmic analysis towards inter and intra superset evaluation and fine-tuning their overlapping coefficients effectively through hybridized mechanism. The convergence time taken to separate the signals was comparatively less when compared to other supervised and unsupervised methods. Results show the proposed scheme consistently reduces background noise with no further apparent speech damage.

Keywords: Speech separation, multichannel, heuristic, K-superset, voiced, unvoiced, binary mask

INTRODUCTION

Speech recognition and identification of users in an environment is a key challenge in devising a proper electronic device to the patients. Hearing-impaired listeners have greater difficulty in understanding speech in the presence of a competing voice. Speech separation is the vital backbone for various speech processing applications such as automatic speech recognition, speaker recognition and audio retrieval and hearing prosthesis. Indoor and Outdoor environment has several factors which further degrades the quality of source speech mixture. Powerful techniques and research works were proposed for speech enhancement in both single channel (Logeshwari and Mala, 2012, 2013) and co-channel. In Adaptive Noise Cancellation (ANC) technique, the process of filtering the noise using filters was adapted while it was not applicable when the interference of speech signal contains another speech signal (Hu and Wang, 2013; Minhas and Gaydecki, 2014). Co-channel speech separation scheme aimed to separate the two clean speech signals transmitted on the same channel using supervised and unsupervised approach (Matheja *et al.,* 2013). The co-channel speech was separated by collecting the speech mixture using two spatially separated microphones. When the speech was accompanied by non-speech noise, the inherent properties of noise was utilized for the speech separation (Shum *et al.,* 2013) but if it was a speech signal, it was not possible to do so. For all the supervised learning

Corresponding Author: Logeshwari, Department of Information and Technology, St. Joseph's College of Engineering, Chennai, India

algorithms, clean utterances should be available apriori for the system to construct speaker dependent models. One of the main drawbacks of supervised learning is the mismatch between training and testing signals. Additionally, existing approaches for multichannel speech separation were mainly based on multiple sources captured using multiple microphones. Simple pitch extraction method was used for multi-speaker speech by identifying the pitch information from temporal processing for spectral processing. Hybrid algorithm considered to have low complexity, computationally efficient and optimized for real-time implementation. Both the information- theoretic and de-correlation approaches were used to achieve superior source separation with fast convergence. Post-separation speech harmonic alignment that resulted in an improved quality of separated speech in a real room environment focusing on separation algorithm for clean speech signals without discussing background noise (Vishnubhotla , 2011). The multichannel adaptive filters were used to remove noise and interference signals using spatial information (Jain and Rai, 2012).

In addition, unvoiced speech poses a big difficulty for multichannel speech separation due to its weak energy and lack of harmonic structure. Therefore, it becomes possible to segregate multichannel speech with a low computational load. Tradeoff between noise reduction and intelligibility is one of the primary issues in speech enhancement. Most of these research were targeted for removing residual noise using supervised mechanism in user speech signals. In these sensitive applications were intelligibility and naturalness are important, non-aggressive setups for speech enhancement algorithms are thus privileged. Objective of the research proposed is to set a more reasonable goal of not affecting the intelligibility (retain its naturalness) of the speech signal in the noise removal process rather than improving it. The approach targets for speech signal processing for multiple users using "unsupervised method" which is the first of its kind. It focus on Multichannel speech separation (Ihara *et al*., 2007; Muhammad, 2012) process. Multichannel speech signals captured in both stationary and non-stationary noisy environment is processed using Multichannel B Hybridized K-Superset Heuristic Speech Separation Algorithm (MC-HKHSA) for efficient speech segregation.

The primary aspect lies in its fine algorithmic analysis towards inter and intra-superset evaluation and fine tuning their overlapping coefficients effectively for same as well as different gender combinations. Objective of the proposed technique is to implement a light-weight computational process (simple and efficient implementation) that aims to, segregate individual speech signals retaining its intelligibility, quality, naturalness, etc. Remove excess background residual noise while retaining the positive features of enhanced speech. The algorithm first splits the given mixture as voiced and unvoiced speech segments. Then from the voiced speech segments, the various pitch values are found out using pitch estimation algorithm. These pitch values are grouped to K-superset for multiple speakers respectively, using a dynamic multi-channel K-superset creation technique. Finally, using hybridized heuristic speech separation process (Pedersen *et al.,* 2007; Reddy and Raj, 2007;Krishnamoorthy and Prasanna, 2010), the voiced and unvoiced segments (the complementary of voiced segment of the remaining persons) of the persons are grouped as a single stream to get the separated speech. The coarse level separation and fine level separation increases the accuracy of the separated speech (Mustafa and Bruce, 2006) making the overall process hybridized. Evaluations demonstrate that our approach outperforms well for more than two speakers with various energy levels.

The improvement was due to our fine algorithmic analysis towards inter and intra superset evaluation and fine tuning their overlapping coefficients effectively. The convergence time taken to separate the signals are less when compared to other supervised and unsupervised methods. The performance of this proposed method was tested using objective quality measures such as percentage of noise residue, the Signal-to-Noise Ratio (SNR) gain, percentage of energy loss and perceptual evaluation of speech quality showed that the MC-HKHS method (Average SNR is 81%) provides an average improvement in the SNR by 8% compared to best performing existing methods (EBSA Average SNR is 73%), respectively (Hidri *et al*., 2012).

MATERIALS AND METHODS

System architecture: Proposed technique follows an unsupervised approach of speech separation where, neither a reference signal nor any prior information regarding the speech or speakers is provided. The algorithm is designed to efficiently separate the speech streams of multiple persons automatically, based on separation criteria that are imposed on the output streams. Its adaptive methodology not only converges faster but is computationally efficient for real-time hardware implementation. Proposed system model for Multichannel Hybridized K-Superset Heuristic Speech Separation System (MC-HKHS) mainly concentrates on unsupervised learning. Proposed MC-HKHSA is a simple, efficient and light-weight computational mechanism for unsupervised multichannel speech separation.

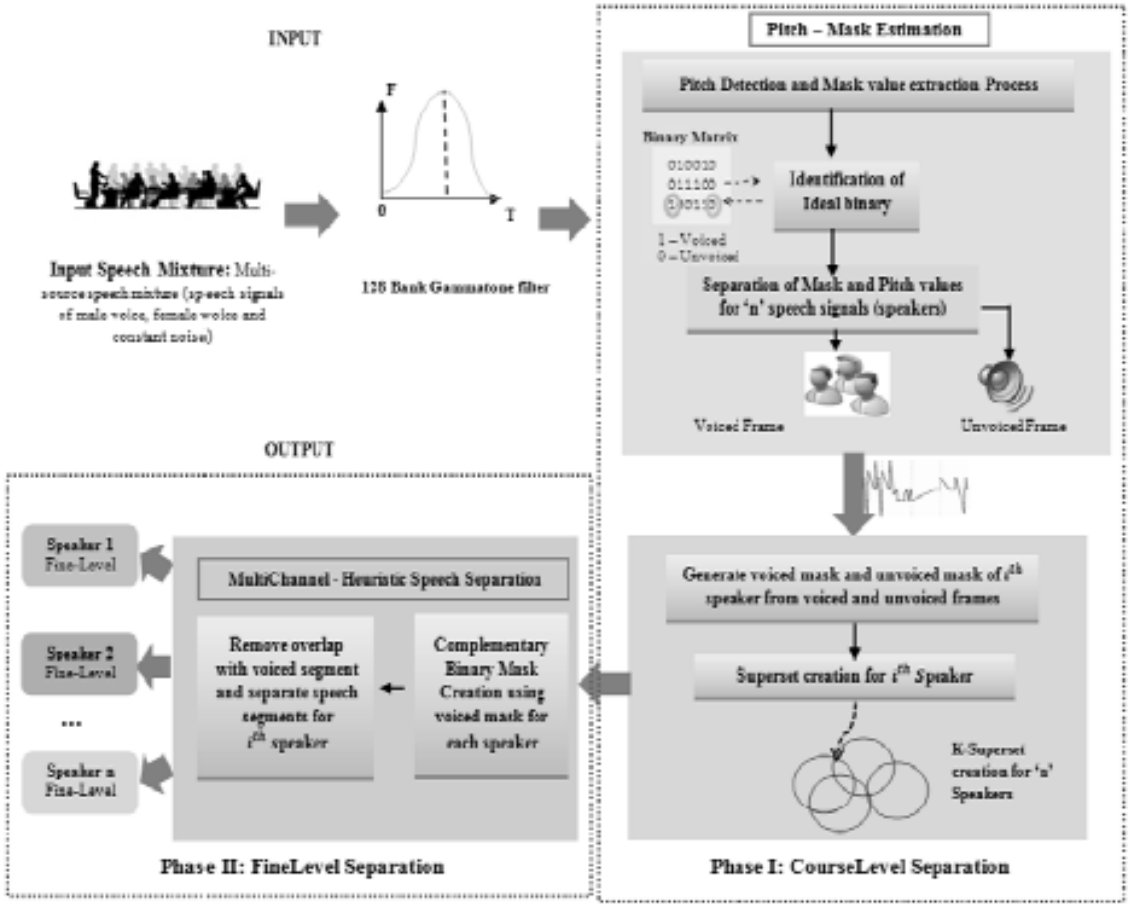

Fig. 1: System model of multichannel-hybridized k-superset heuristic speech separation system

Table 1: Nomenclature used in MC-HKHSA scheme

P_{est}	Pitch estimation algorithm
G_f	Gammatone filter
$N_{Channel}$	Number of channels
S_p^i	ith speaker
P_i	ith pitch
F_r^i	ith frame
UV^i_{mask}	i^t unvoiced mask
V^i_{mask}	i^t voice mask
U_{mask}	Unvoiced mask
$V^i_{cluster}$l	i^{t}Vector cluster
$\vec{P}$	1D vector of pitch values
$K_{superset}$	Number of supersets
M^i_{sk}	ith mask
X_{mat}	Speech mixture
N_{frame}	Number of frames
C^i_{mask}	i'complementary binary mask
IBM^i_{mask}	ith IBM mask
E^i_{frame}	Energy of i Frame

Overall system model of MC-HKHS is depicted in Figure 1 and nomenclature used in the proposed scheme is referred in Table 1. Proposed MC-HKHS System is categorized into two phases:

- Coarse level separation
- Fine level separation

Coarse level separation: Mixture (X captured in noisy environment using single microphone for 60 sec duration considered as the input. The speech mi down sampled using 128 channel Gammatone filter bank for quick processing. As the frequency components in the spectrum of the speech signal are about 7.5 kHz, the sampling frequency is chosen as 16 kHz speech mixture is divided into frames of length 20 ms and overlapping by 10 ms. voiced mask and unvoiced mask is generated from the voiced and unvoiced frame input speech mixture. i.e., for each and every frame, the voiced and unvoiced portions are identified. The binary mask set as 1 for voiced speech (V_{mask})and 0for unvoiced speech (U_{mask}). Pitch value are extracted from the voiced speech segments. The pitch values will be 0 for all the unvoiced frames Eq. 1.

$$(V_{mask}, U_{mask}, Pitch_{value}) = mask\_pitch\_estimation(Speech_{mixture}) \quad (1)$$

These pitch values along with the (V^i_{mask}) are grouped into multiple (K) cluster speakers respectively. Grouping into K clusters is achieved using a dynamic heuristic clustering algorithm with the help of Silhouette value called Hybrid Vector Quantization. This algorithm estimates the number of different range of pitch values in the given mixture. Silhouette refers to a method of interpretation and validation of clusters of data. If the silhouette value is about zero, it means that this entity could be assigned to another cluster as well. If the value is close to -1 then it refers that the entity is misclassified and if all the silhouette width values are close to 1, means that the set is well clustered. Therefore the threshold value for a silhouette is set to 0.7, so that a good set of clusters are formed.

$$V^i_{mask} = \text{cluster}\left(V_{mask}\right) \tag{2}$$

$$UV^i_{mask} = \text{setdiff}\left(V_{mask}, V^i_{mask}\right) \tag{3}$$

The coarse level separation is done with the help of mask values and cluster values for each different speech signal. Steps that belong to coarse level separation is elaborated as referred as:

- Obtain a rough estimate of target pitch
- Segregate target speech using harmonicity and temporal continuity
- Extract majority of target speech without including much interference
- Identify the Ideal Binary Mask (IBM) with a Time–Frequency (T-F) representation The IBM is a binary matrix along time and frequency where 1 indicates that the target is stronger than interference in the corresponding T-F unit and 0 otherwise
- The input signal is then decomposed in the frequency domain with a bank of 128 gammatone filters (G_f with their center frequencies equally distributed on the equivalent rectangular bandwidth rate scale from 50-8000 Hz
- In each filter channel, the output is divided into 20-ms time frames with 10ms overlap between consecutive frames

$$A(c,m,T) = \frac{\sum_n x\left(c, mT_m - nT_n\right) - x\left(c, mT_m - nT_n - TT_n\right)}{\sqrt{\sum_n x^2\left(c, mT_m - nT_n\right)}} \tag{4}$$

Using cross-channel correlation measure the similarity between the responses of two adjacent filters. Indicate whether the filters are responding to the same sound component. Calculate the cross-channel correlation by using Eq. 5 represented as:

$$A(c,m) = \frac{\sum_T \left[A(c,m,T) - A\left(\widehat{c,m}\right)\right]\left[A(c+1,m,T) - A\left(\widehat{c+1,m}\right)\right]}{\sum_T \left[A(c,m,T) - A\left(\widehat{c,m}\right)\right]^2 \sum_T \left[A(c+1,m,T) - A\left(\widehat{c+1,m}\right)\right]^2} \tag{5}$$

T-F unit is labelled 1 if the corresponding response has a periodicity similar to that of the target. Let H_0 be the hypothesis that a T-F unit is target dominant and H_1 otherwise. The instantaneous frequency of the response within a T-F unit is simply as half the inverse of the interval between zero-crossings of the response. When interference contains one or several harmonic signals, there are time frames where both target and interference are pitched. In such a situation, it is more reliable to label a T-F unit by comparing the period of the signal within the unit with both the target pitch period and the interference pitch period. In particular, Tcm should be labelled as target if the target period not only matches the period of the signal but also matches better than the interference period Eq. 6:

$$\begin{aligned} &p\left(H_0 \middle| T_{cm}\left(T_s(m)\right)\right) > \\ &p\left(H_0 \middle| T_{cm}\left(T'_s(m)\right)\right) p\left(H_0 \middle| T_{cm}\left(T_s(m)\right)\right) > 0.5 \end{aligned} \tag{6}$$

Where, T_s' is the pitch period of the interfering sound at frame m. Equation 6 is used to label T-F units for all the mixtures of two utterances in the test corpus. A better performance isobtained by using the pitch values of speakers. Labelling a T-F unit using only the local information within the unit still produces a significant amount of error. Since speech signal is wide band and exhibits temporal continuity, neighbouring T-F units potentially provide useful information for unit labelling obtained by using the pitch values of multiple F unit using only the unit still, produces a significant amount of error. Since, speech signal is wide band and exhibits temporal F units potentially provide useful information for unit labelling where, information from a neighbourhood of T-F Units is also considered. From the estimated mask of the voiced target, target pitch is to be estimated. Let {c} be the set of binary mask labels at frame m where L(c,m) is 1 if is active and 0 otherwise.

A frequently used method for pitch determination is to pool autocorrelations across all the channels and then identify a dominant peak in the summary correlogram. As

Speech signals exhibit temporal continuity, the pitch and the ideal binary mask of a target utterance tend to have good temporal continuity. This concept is used to further A frequently used method for pitch determination is to pool autocorrelations across all the channels and then identify a dominant peak in the summary correlogram. As Speech signals exhibit temporal continuity, al binary mask of a target utterance tend to have good temporal continuity. This concept is used to furtherimprove pitch estimation. The reliability of the estimated pitch based on temporal continuity is verified and validated, for every three consecutive frames, m-1, m, m+1 pitch changes are all <20%.

$$\begin{aligned}&\left(T_s(m)\right)-\left(T_s(m-1)\right)<0.2\min(T_s(m))\\&(T_s(m-1))\left|(T_s(m))-(T_s(m+1))\right.\\&0.2\min(T_s(m)),(T_s(m+1))\\&\sum_c L(c,m)L(c,m-1>)\\&\sum_c L(c,m)L(c,m+1)\end{aligned} \tag{7}$$

The estimated pitch periods in these three frames are all considered reliable. The unreliable pitch points are re-estimated by limiting the plausible pitch range using neighbouring reliable pitch points for two consecutive time frames, m-1 and m. Then an initial estimate of pitch periods are generated in each time frame as pitch contours and binary masks for up to three sources; a pitch contour refers to a consecutive set of pitches that is considered to be produced by the same sound source. Then the estimation of pitch contours and masks in an iterative manner is improved. Let $T_{S,1}(m)$, TS,2(m) and $T_{S,3}(m)$ represents three estimated pitch periods at frame m. $L_1(m)$, $L_2(m)$ and $L_3(m)$ are the corresponding labels of the estimated masks. With the estimated pitch period $T_{S,1}(m)$, we re-estimate the mask as:

$$\begin{aligned}&L_1(c,m)\\&\quad 1,p(H_o\left|T_{cm}(T_{s,1}(m))\right.)\le\theta_p\text{and}\\&C(c,m)>0.9850rC_E(c,m)>0.985\\&=\{0,\text{else}\end{aligned}$$

$$\begin{aligned}&L_2(c,m)\\&\quad 1,p(H_o\left|T_{cm}(T_{s,2}(m))\right.)\le\theta_p\text{and}\\&C(c,m)>0.9850rC_E(c,m)>0.985\\&=\{0,\text{else}\end{aligned}$$

$$\begin{aligned}&L_3(c,m)\\&\quad 1,p(H_o\left|T_{cm}(T_{s,3}(m))\right.)\le\theta_p\text{and}\\&C(c,m)>0.9850rC_E(c,m)>0.985\\&=\{0,\text{else}\end{aligned}$$

Through iterative mechanism, majority of the target voiced speech is extracted without any interference. Finally, the extracted segments of same pitch (pitch refers to the consecutive set of voices that belongs to the same source) are combined into single frame to produce good temporal continuity for speech signals. The output is a set of simultaneous streams (through binary masking and their associated pitch contours) in which voiced speech is represented as 1 and unvoiced speech is represented as 0. The same procedure is repeated for all the clusters (here in after referred as "supersets") to separate the speech segments for 'K' speakers. This pitch estimation and voice speech segregation is aneffective and robust process and tend to produces good estimates of both pitch and voiced speech even in the presence of strong interference. Figure 2 shows the flow model for coarse level speech separation of speech mixture. In order to further fine tune the speech supersets, multiChannel-heuristic speech separation technique is performed.

Fine level separation using multichannel heuristic speech separation: The fine level separation further increases the accuracy of the separated speech signal. In fine level Separation, the unvoiced speech segregation is done by generating unvoiced T-F segments using bi-polar coefficient detection based segmentation. Then group unvoiced segments in unvoiced-voiced intervals using the complimentary mask of the segregated voiced speech. The overlap between an unvoiced segment and the complementary binary mask of segregated voiced speech is calculated for each speaker and the segment (superset), is assigned accordingly. Each K-Superset has 'n' frames, i.e., The target speech (including voiced and unvoiced speech) and interference are correctly segmented and are well separated into different segments using multiscale bi-polar coefficient detection based analysis. Bi-polar coefficient is a pair of value where the peak value denotes to a sudden increase of acoustic energy and is start of auditory events and valleys denotes the ends of corresponding events. Segments are then produced by pairing bi-polar coefficient fronts in multiple scales. Since bi-polar coefficient detection based segmentation utilizes energy fluctuations, the segments thus formed include both voiced and unvoiced speech. To retain only unvoiced segments, the parts of segments overlapping with segregated voiced speech are removed. , i.e., any T-F

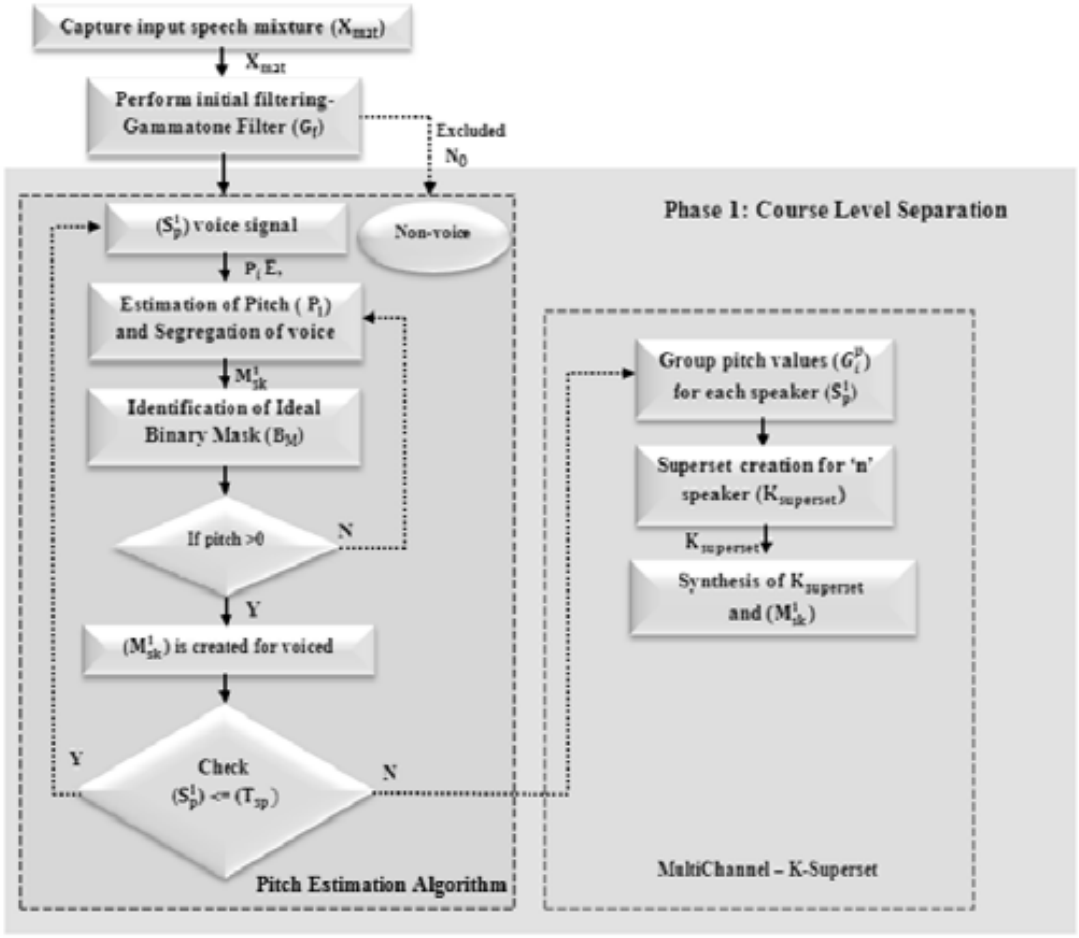

Fig. 2: Flow process of coarse level of speech mixture

unit in bi-polar coefficient detection based segments (and also included in segregated voiced speech) is removed.

Figure 3 shows the flow model for fine level speech separation of speech mixture. Contiguous T-F regions in the remaining parts thus correspond to unvoiced segments. input mixture is passed through a bank of gammatone filters. To extract its temporal envelope, the output from each filter channel is half-wave rectified, low-pass filtered (a filter with a 74.5 m sec Kaiser window and a transition band from 30-60 Hz) and down sampled to 400 Hz. The temporal envelope, indicating the intensity of a filter output is used for bi-polar coefficient detect analysis. Smoothing corresponds to low filtering. The system further smoothes the intensity over frequency with a Gaussian kernel, to enhance the alignment of bi coefficients. The degree of smoothing is high if the scale is larger. At a certain scale, bi polar coefficient candidates are detected by marking peaks and valleys of the time derivative of the smoothed intensity into one segment.

Multiscale integration, integrates analysis at different scales to form segments. The system captures a majority of speech events at the largest scale but misses some small segments. As the system integrates analysis at smaller scales more speech segments are formed; at the same time more segments from interference also appear.

It locates more accurate bi-polar coefficient positions for segments and new segments can created within the current background. Thus the voiced and unvoiced speech and interference are correctly segmented and are well separated into different segments. Then group unvoiced segments in unvoiced-voiced intervals using the complimentary mask of the segregated voiced speech. Complementary binary mask (C^i_{mask}) is created for each speaker with the help of a voiced mask (V^i_{mask}) of the corresponding speaker and a complement of voiced mask for the remaining speakers, i.e., create a IBM $^i_{mask}$ with 1 as the value for the entire mask and U V^i_{mask} with 0 as the value for the entire mask, for each user. Assign 0 to the

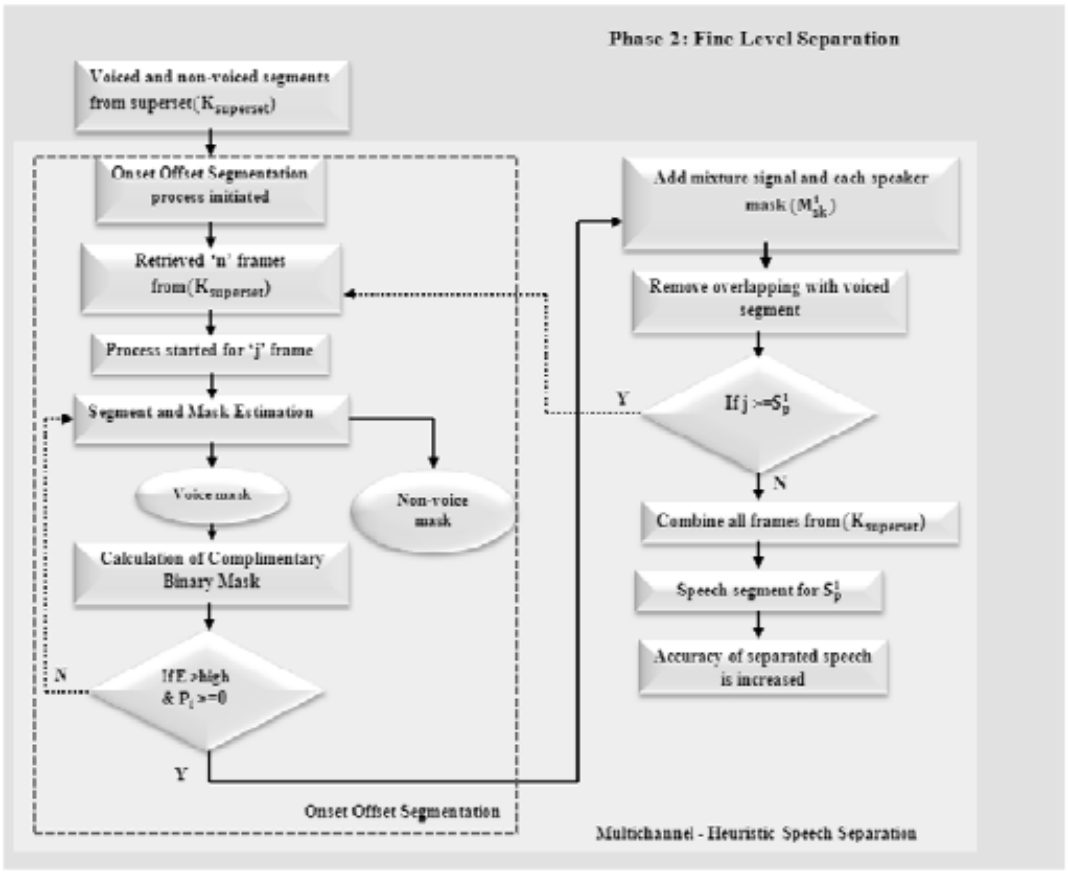

Fig. 3: Flow process of fine level separation of speech mixture

segments where pitch value is zero in voiced mask (V^i_{mask}) of that speaker. Calculate the complimentary binary mask (CBM) using the formula, Eq. 8:

$$C^i_{mask} = (1 - V^i_{mask}) * M^i_{sk} * (1 - M^{(i+1)modn}_{sk}) \quad M^{(i+2)modo}_{sk} \tag{8}$$

If the sum of the energy level multiplied by the complimentary mask of speaker 'i' is greater than the sum of the energy level multiplied by the complimentary mask of remaining speakers, then assign 0 to U V^i_{mask} for all the frames whose $Pitch_{value} \geq 0$ for speaker 'i'. For the unvoiced segments (overlapping in time with the voiced speech of a segregated speaker) group based on the already-segregated voiced speech. Add the entire mask and remove the overlap if any between this voiced binary mask created for each user and the unvoiced binary mask Eq. 9:

$$M^i_{sk} = remove - overlap(V^i_{mask} + UV^i_{mask} + U_{mask}) \tag{9}$$

Unsupervised segregation of unvoiced portions is extremely challenging. Such portions, however is split equally to all speaker's group. Speech segments are separated for that corresponding speaker 'I' as given in Eq. 5. Inverse FFT, synthesis of input mixture with the generated binary mask 'i' for each speaker is performed. Thus, all simultaneous streams are grouped into K-supersets (speech streams) where eachsuperset corresponds to the voiced speech of one speaker Eq. 10.

$$S_i = IFFT(M^i_{sk} * Speech_{mixture}) \tag{10}$$

Thus the accuracy and efficiency of the separated speech is increased in the fine level separation process. For segments in unvoiced-unvoiced intervals, we separate them by a simple split. Lastly, our system combines the estimated voiced and unvoiced masks to form 'K' complete speaker masks. A reasonable estimate of number of speakers was obtained by adapting the use of a heuristic speech separation for K-superset creation. The algorithmic process for Multichannel heuristic speech separation process is referred as.

No ofmutichannel-heuriistics speech separation process:
Recive voiced and non-voiced segment and segment 'n'frames from 'k'supersets
N_{subset}*N_{frame}*$N_{channel}$;
Voiced segment and mask estimation:
$M^i_{sk} = V^i_{mask}$+ U V^i_{mask}+U_{mask}
Complimenttary Binary Mask (CBM):
voiced speech op i^{th} speaker;
$C^i_{mask} = (1-V^i_{mask})*M^i_{sk}*(1-M_{sk}(i+1) \bmod_n - M_{sk}(i+2) \bmod_n)$
Calculate energy level for each frame and create a binary mask;
For j = 1: E^i_{frame}
If (E^i_{frame}>high) and (Pitch (P_i))> = 0 then
"Add mixture signal and mask (voice and non-Voiced)"
$M^i_{sk} = M^1_{sk}+M^2_{sk}+...+M^n_{sk}$
If (j> = sp^i)
Combine all the frames from K-superset
End
End
$M^i_{sk} = V^i_{mask}$+ U V^i_{mask}+U_{mask}
Remove overlap with voiced segment
M^i_{sk} = remove_overlap (V^i_{mask}+ U V^i_{mask}+U_{mask})
Speech segment is separated for thei^{th} speaker;
Accuracy of separated speech is increased;
s_i = IFFT (M^i_{sk} × $speech_{mixture}$)

If the number of speakers needs to be estimated, improved results are obtained via an iterative optimization procedure which alternates between K-supersets and re- segmentation until the separation hypothesis converges. The iterative optimization procedure gives the system more opportunities to re-estimate the number of speakers using (hopefully) cleaner and more refined speech segments. For noisy environments, the speech mixture is filtered using filters like Weiner filter and Spectral subtraction to remove the stationary and non stationary noises and the filtered speech mixture is processed in the same manner as clean speech mixture.

RESULTS AND DISCUSSION

Simulation and experimental analysis: Our Multichannel Hybridized K-Superset Heuristic Speech Separation System (MC-HKHSS) was implemented in MATLAB interfacing with native C for Pitch and mask evaluation and bi-polar detection based segmentation. Speech samples were taken from grid corpus and also, tested with self recorded samples. Experimental analysis was conducted by capturing the signal in the noisy environment, with various noise levels ranging between 50-145 db. Internal and external noise were considered in this experimental process. The noisy signals are generated by adding noise to the clean speech (test sentences). Types of noise considered are vehicle, i.e., motorbike noise (common mode of transport in most Indian cities); fan, A/C noise (typical office environment with a ceiling fan); and car noise. These noise types were recorded separately. We conduct the experiments for SNR: -10, -5, 0, 5 and 10 dB. During the experimental analysis, speech mixture for 3 users was considered. During simulation MC-HKHSS derived 3-Supersets at coarse Level separation process as depicted in Fig. 4. For the sake of clarity and concision, the output scores were averaged over all types of noise for each of the SNR conditions. For each SNR condition, we randomly created 50multichannel speech mixtures for testing from the grid database. Among them, 15 are male-male-female mixtures, 10 are male mixtures, 10 are female mixtures and 15 are female-female-male mixtures. All test mixtures are down sampled from 25 kHz to 16 kHz for faster processing.

The performance of the system is evaluated based on the SNR gain of the Target Signal. The` SNR gain is calculated as the output SNR of separated speech subtracted by the input SNR. For each separated speech, we take the synthesized speech from the overall IBM as the ground truth and measure the output SNR as,

$$SNR_{value} = 10 \log_{10} \left[\frac{\sum_i (s(t))^2}{\sum_i s(t) - s'(t)^2} \right]$$

Where
s(t) = Original speech signal
s'(t) = Extraacted speech signal

Table 2 and 3 shown below displays the average SNR for 10 mixtures at 10, 5, 0, -5 and -10 db using MC-HKHSA (clean mixture) for 3 speakers. Figure 5 and 6 shown below displays the average SNR for 10 mixtures at 10, 5, 0, -5 and -10db using MC-HKHSA for clean and noisy mixture for 3 speakers.

Figure 7 displays the speech mixture with noise considered during simulation process. Adaptive noise cancellation using filtering technique (Weiner filter and Spectral subtraction was used to remove the stationary noise) resulted in enhanced speech separation. Verification at various levels was performed during the experimental process to estimate the extent to which MC-HKHS System was capable of identifying and removing such stationary and non-stationary noises misture. Figure 8 shows the noise signal separated from speech.

It proved that MC-HKHS System efficiently removed the entire stationary noise (like fan or AC noise) that existed in entire speech and also in silence or clean period using spectral subtraction algorithm, resulting in

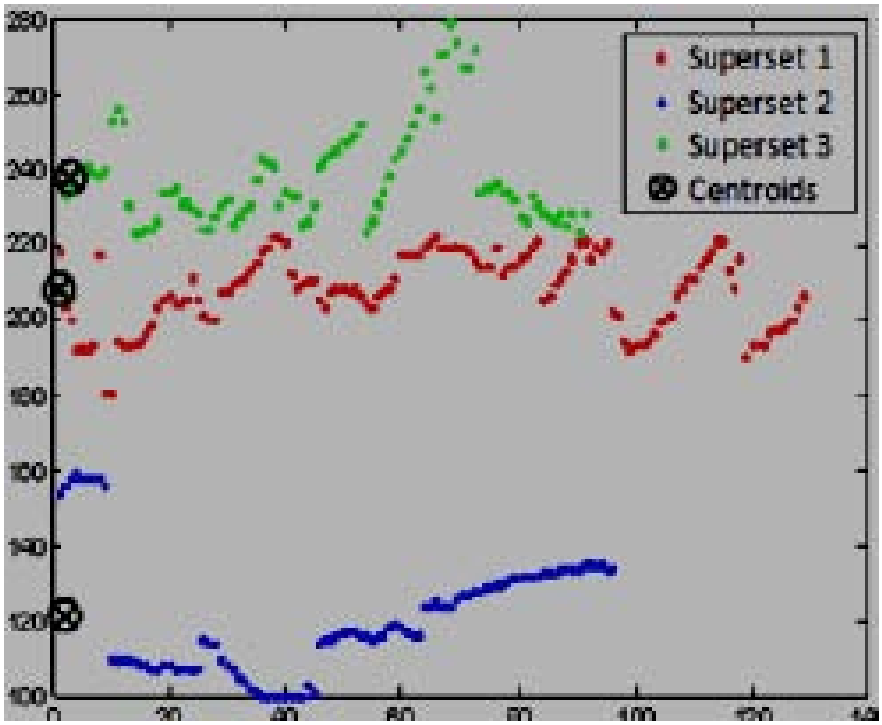

Fig. 4: MC-HKHSS derived 3-supersets at coarse Level separation process

Table 2: Average SNR for 10 mixtures at 10, 5, 0, -5 and -10 db using MC-HKHSA (noisy mixture)-3 speakers

Input	Gender	-10	-5	0	5	10
M1+M2+F1 mixture	Male 1	14.2645	14.4119	14.5409	14.3528	14.1833
	Male 2	14.2116	14.4936	14.7181	14.551	14.1237
	Female 1	15.0917	14.6292	14.4736	14.4717	14.6014
M1+F1+F2 mixture	Male 1	14.6301	14.3392	14.571	14.2995	14.0478
	Female 1	14.7791	14.3299	14.468	14.1234	14.4127
	Female 2	14.9351	14.4025	14.3866	14.1337	14.1949
F1+F2+F3 mixture	Female 1	14.2178	14.3971	14.5273	14.4298	14.1064
	Female 2	14.1835	14.3413	14.4458	14.3986	14.0296
	Female 3	14.1307	14.4281	14.5019	14.1424	14.1827
M1+M2+M3 mixture	Male 1	14.6392	14.5219	14.7857	14.5376	14.1848
	Male 2	14.9998	14.7943	14.5411	14.2046	14.0703
	Male 3	14.8802	14.3109	14.4339	14.3177	14.0285

Table 3: Average SNR for 12 mixtures at 10, 0, -5 and -10 db using MC-HKHSA (Clean mixture-3 speakers

Input	Gender	-10	-5	0	5	10
M1+M2+F1 mixture	Male 1	16.8076	16.1232	16.6189	16.9265	16.5781
	Male 2	16.4171	16.5634	16.7171	16.6104	16.3717
	Female 1	16.1264	16.2772	16.3596	16.0742	16.0265
M1+F1+F2 mixture	Male 1	16.4584	16.3827	16.1948	16.0729	16.0162
	Female 1	16.3173	16.2279	16.1676	16.0323	16.9117
	Female 2	16.2672	16.1308	16.0204	16.8706	16.5412
F1+F2+F3 mixture	Female 1	16.3925	16.2183	16.4374	16.1591	16.0436
	Female 2	16.3255	16.1182	16.3016	16.1163	16.9172
	Female 3	16.7425	16.9631	16.0826	16.8271	16.6615
M1+M2+M3 mixture	Male 1	16.4718	16.437	16.6721	16.4261	16.3165
	Male 2	16.5423	16.7254	16.8166	16.6182	16.3472
	Male 3	16.6194	16.2173	16.6229	16.9548	16.8193

Figure 9 effectively separating the speech signals from the mixture. The mixture of multiuser speech signal after excluding noise. The unsupervised approach for multi channel speech separation is designed using Hybrid Vector Quantization Heuristic Clustering Algorithm. The algorithm gives an accuracy

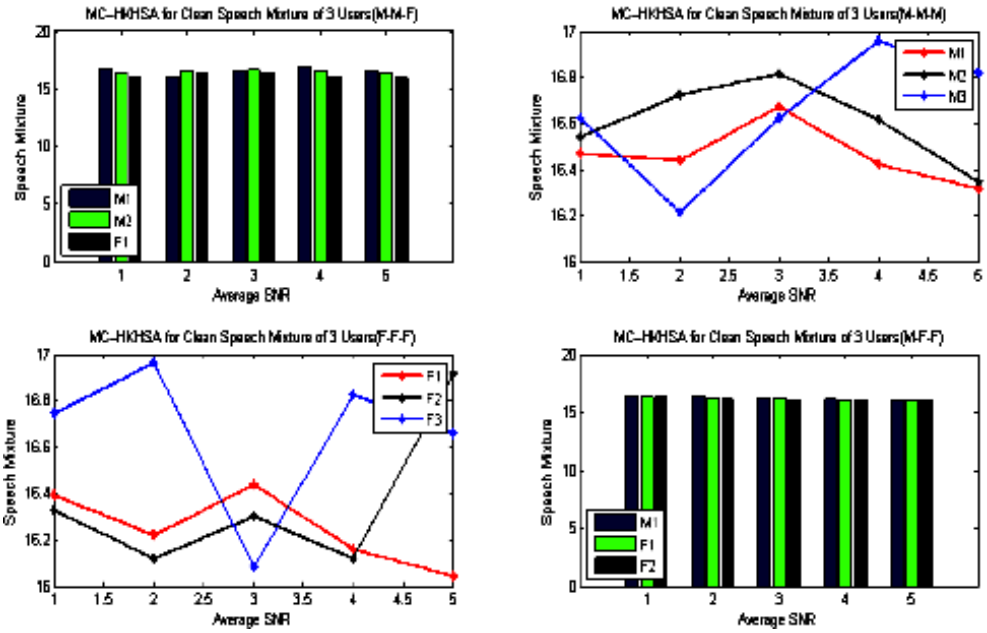

Fig. 5: Average SNR for 12 mixtures at 10, 5, 0, -5 and -10 db using MC-HKHSA (Clean mixture - 3 speakers)

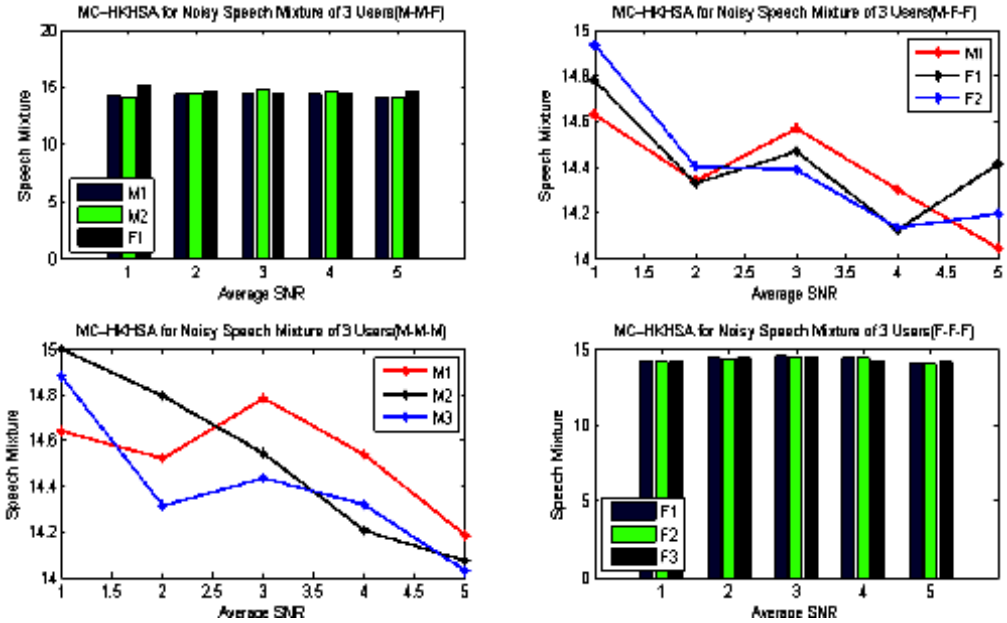

Fig. 6: Average SNR for 10 mixtures at 10, 5, 0, -5 and -10 db using MC-HKHSA (Noisy mixture - 3 speakers)

of about 81% for clean mixture with two and three speech mixtures.

For same gender mixtures, the algorithm works from two to four different speechs. For different gender mixtures, the algorithm research from 2-8 different speech signals Figure 10-12 shows the speech signals of all 3 users after speech separation processing. Figure 13 shows the spectrogram mixture of speech signals of all 3 users followed by Fig. 14 displaying the spectrogram of separated speech signals of individual users.

We compare our system with other supervised and unsupervised methods across a range of input SNR conditions. Figure 15 illustrates the comparison between MC-HKHSA (3 speakers) for speech separation in clean

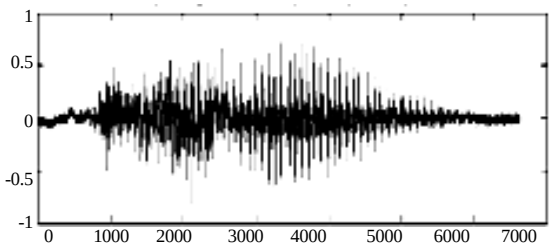

Fig. 7: Mixture of 3 speecd signals with noise

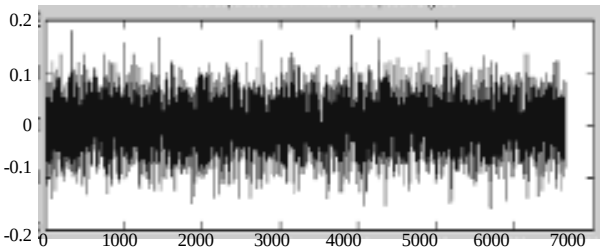

Fig. 8: Noise signal separated from speech mixture

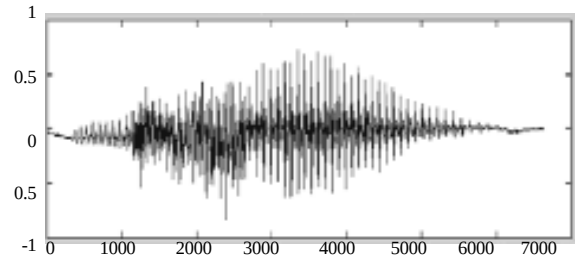

Fig. 9: Mixture of 3 speecd signals excluding noise

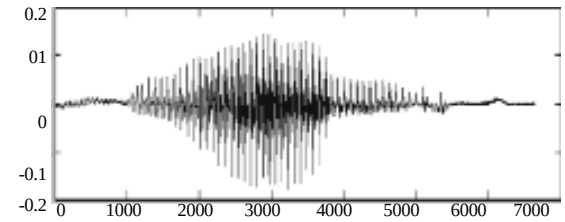

Fig. 10: Separated speech signal of speaker 1

and noisy environment. From the results we infer that, MC-HKHS algorithm outperforms existing methods (EBSA) in terms of average SNR both in a clean and noisy

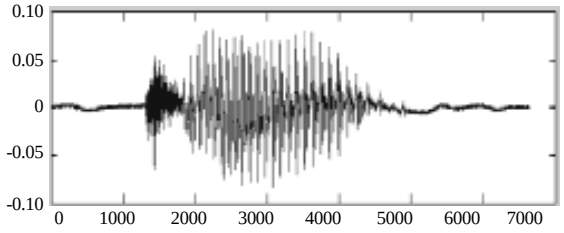

Fig. 11: Separated speech signal of speaker 2

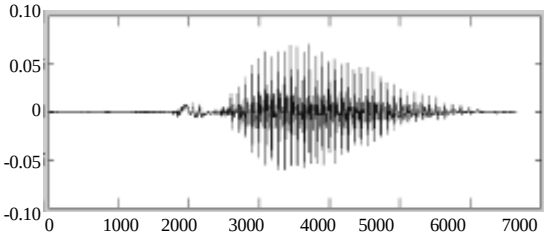

Fig. 12: Separated speech signal of speaker 3

environment. Average SNR value of MC-HKHS Algorithmis 81% when compared to EBSA's average SNR value which is 73% in clean and noisy environment. Evaluations demonstrate that the MC-HKHSA procedure converges quickly in a considerable range of SNRs and improves separation results significantly. The input SNR is then used to adapt the speaker models for more accurate estimation. This factor of improvement was due to our fine algorithmic analysis towards inter and intra supersetevaluation and fine tuning their overlapping coefficients effectively.

Performance Graph of Male-Male-Female (MMF), Male-Female-Female (MFF), Female-Female-Female (FFF) and Male-Male-Male (MMM) Mixtures using MC-HKHSA for 3 speakers is shown in Fig. 16. This performance evaluation depicts that MC-HKHSA is clearly a multichannel speech separation algorithm for 'n' number of users (multi-gender). It was found that the convergence time taken to separate the signals are less when compared to other supervised and unsupervised methods. Our future scopeof work will also involve human voice/speech along with animals, birds sounds and other sounds very specific to an environment like railway stations, bus terminus, airports, markets, public gathering, etc., Overall performance of MC-HKHS Algorithm is depicted in comparison graph Fig. 16, for a clear understanding to the readers and reviewers of the scientific community.

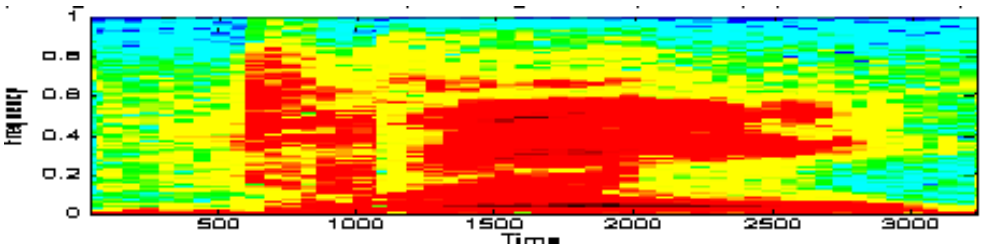

Fig. 13: Spectrogram mixture of 3 users

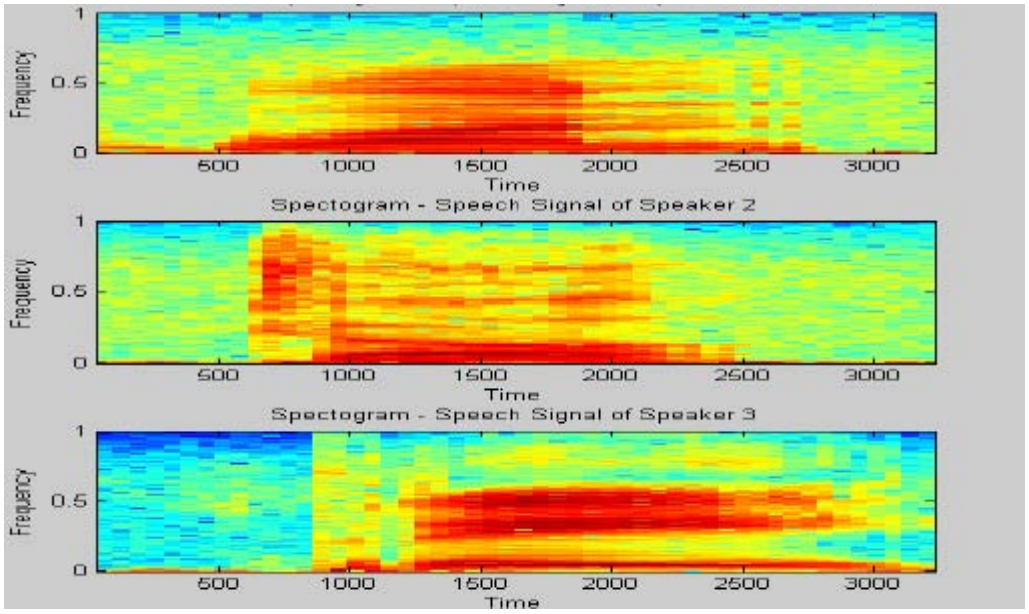

Fig. 14: Spectrogram of separated speech signals of 3 users

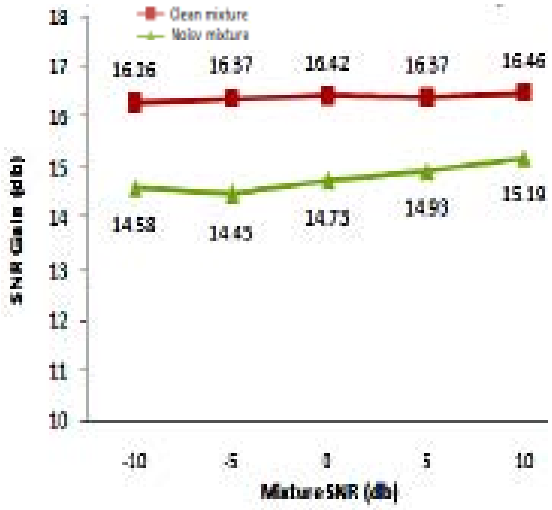

Fig. 15: Multichannel - HKHSA for speech separation in noisy and clean environments-3 speakers

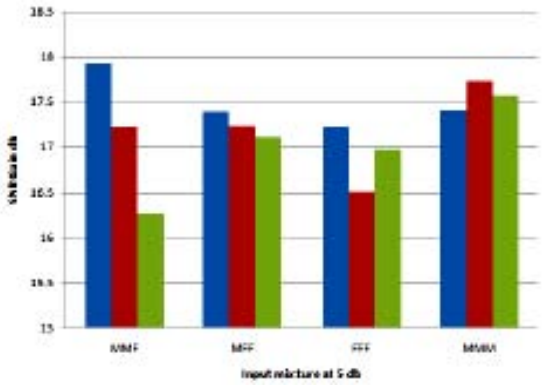

Fig. 16: Performance graph of MMF, MFF and FFF and MMM mixture using MC- HKHSA for 3 speakers

CONCLUSION

A novel unsupervised approach to multichannel speech separation is proposed using MC-HKHS Algorithm in

both stationary and non-stationary noisy environment. Noise elements in the input mixture are filtered using Weiner and SS filter. Then P EA is used for pitch estimation, proposed multichannel K-superset creation technique was used for grouping simultaneous streams across time and separating them to a coarse level, followed by multichannel heuristic speech separation mechanism that involves bi-polar coefficient detection based segmentation technique to separate the signal to a fine level andmeasures the speaker difference of each hypothesized K-superset and incorporates pitch constraints to generate the fine speech. The calling bell sound, phone bell sound which has pitch can also be eliminated from the input mixture. Systematic evaluations and comparisons show that our method achieves considerable SNR gains. Despite its unsupervised nature, it produces comparable performance to model-based and speaker independent methods. In this research, our K-superset creation algorithm is derived for multi channel speech with four speakers. The algorithm can be extended to deal with number of speakers. Our system may refined further more considering various types of external and intenal noises. To the best of our knowledge, this approach is one of the novel and true hybridized mechanism involving simultaneous process management, thusreducing the error rate due to inter algorithmic value conversions. Future scope of this rechearch will also involve human voice/speech along with animals, birds sounds and other sounds very specific to an environment like railway stations, bus terminus, airports, markets, public gathering, etc., for guiding elder/aged people having hearing impaired problems and for sophisticated voice recognition systems. This proposed system would be able to separate multiple speech (K = “n” users), in a mixture under a strict conditions that no more than one users have the same fundamental frequency and pitch value. This is one of the limitations in unsupervised methods and can be solved only using known supervisory techniques.

REFERENCES

Hidri, A., S. Meddeb and H. Amiri, 2012. About multichannel speech signal extraction and separation techniques. J. Signal Inf. Process., 3: 238-247.

Hu, K. and D. Wang, 2013. An iterative model-based approach to cochannel speech separation. EURASIP. J. Audio Speech Music Process., 2013: 1-11.

Ihara, T., M. Handa, T. Nagai and A. Kurematsu, 2007. Multichannel speech separation and localization by frequency assignment. Electron. Commun. Japan Part III Fundam. Electron. Sci., 90: 59-70.

Jain, S.N. and C. Rai, 2012. Blind source separation and ICA techniques: A review. Int. J. Eng. Sci. Technol., 4: 1490-1503.

Krishnamoorthy, P. and S.M. Prasanna, 2010. Two speaker speech separation by LP residual weighting and harmonics enhancement. Int. J. Speech Technol., 13: 117-139.

Logeshwari, G. and G.A. Mala, 2012. A Survey on Single Channel Speech Separation. In: Advances in Communication, Network and Computing. Vinu, V.D. and J. Stephen (Eds.). Springer Berlin Heidelberg, Berlin, Germany, ISBN: 978-3-642-35614-8, pp: 387-393.

Logeshwari, G. and G.A. Mala, 2013. An efficient speaker recognition system for separating the single channel speech using frequency modulation. Int. Rev. Comput. Software IRECOS., 8: 632-641.

Matheja, T., M. Buck and T. Fingscheidt, 2013. A dynamic multi-channel speech enhancement system for distributed microphones in a car environment. EURASIP. J. Adv. Signal Process., 2013: 1-21.

Minhas, S.F. and P. Gaydecki, 2014. A hybrid algorithm for blind source separation of a convolutive mixture of three speech sources. EURASIP. J. Adv. Signal Process., Vol. 2014, 10.1186/1687-6180-2014-92

Muhammad, A., 2012. Multichannel wiener filtering for speech enhancement in modulation domain. Master Thesis, School of Engineering, Blekinge Institute of Technology, Karlskrona, Sweden.

Mustafa, K. and I.C. Bruce, 2006. Robust formant tracking for continuous speech with speaker variability. IEEE. Trans. Audio Speech Lang. Process., 14: 435-444.

Pedersen, M.S., J. Larsen, U. Kjems and L.C. Parra, 2007. A Survey of Convolutive Blind Source Separation Methods. In: Springer Handbook on Speech Processing and Speech Communication, Benesty, J., M.M. Sondhi and Y. Huang (Eds.). Springer, Berlin, Germany, pp: 1-34.

Reddy, A.M. and B. Raj, 2007. Soft mask methods for single-channel speaker separation. IEEE. Trans. Audio Speech Lang. Process., 15: 1766-1776.

Shum, S.H., N. Dehak, R. Dehak and J.R. Glass, 2013. Unsupervised methods for speaker diarization: An integrated and iterative approach. IEEE. Trans. Audio Speech Lang. Process., 21: 2015-2028.

Vishnubhotla, S., 2011. Segregation of speech signals in noisy environments. PhD Thesis, University of Maryland, College Park, Maryland. http://drum.lib.umd.edu/handle/1903/11525.