Neural, Parallel, and Scientific Computations 24 (2016) 489-504

SINGLE CHANNEL SPEECH SEPARATION BASED ON PCA AND FUZZY LOGIC

BELHEDI WIEM, BEN MESSAOUD MOHAMED ANOUAR, AND BOUZID AICHA

Université de Tunis El Manar, École Nationale d'Ingenieurs de Tunis
Laboratoire du Signal, Images et Technologies de l'Information, BP 37
Le Belvédère 1002, Tunis, Tunisie

ABSTRACT. In this paper we propose a new approach of monaural speech separation in an unsupervised manner. This approach is based on the amplitude and phase spectrum of principal component analysis (PCA). It consists in using sparse matrix decomposition and low-rank technique in the spectral domain. This technique is able to distinguish a variable foreground from a relatively more regular background. The proposed method decomposes the composite matrix into two sub-matrices, and then the masking technique is applied on the real part of each subspace giving three different separated signals in three channels. The channel classification is done by means of the Fuzzy logic; it is mainly based on our transformation of the multi-scale product analysis. Importantly, the proposed approach requires only the mixture signal as input with no prior knowledge about the mixing process or about the desired speaker and it works independently from pitch. We evaluate our approach on the Cooke database. The performance results are compared to other studies. Experimental results refer its satisfying effectiveness compared to the state-of-the-art methods.

Key Words: Monaural speech separation, Subspace decomposition, Binary mask, Fuzzy logic, Multi-scale product.

1. Introduction

The sound is important to human beings because of its contents. However, the voice is a spoken language which the human brain can, normally, recognize and deal with naturally but it is a challenge to make a machine execute the same task. This problem received a lot of attention but it has always been negotiable. The most difficult case of separation remains that of mono-channel case because there is no mutual relationship between the channels. In a real environment the mono-channel case is the only case to be treated and that is what motivated us to move towards it. While monaural speech segregation by machines remains a great challenge, several mathematical and technical tools have been developed to help researchers to perform this task.

In literature, the methods that were used for speech separation are divided into three main categories: the first one is based on computational auditory scene analysis

Received May 7, 2016

(CASA) [1, 2, 3], the second one uses sinusoidal modeling [4, 5, 6] and the last is subspaces decomposition methods [7, 8, 9, 10].

A CASA system generally follows four steps: analysis device, the extraction of its properties, the segmentation and grouping. The peripheral processing decomposes the auditory scene representation in a two-dimensional time-frequency (TF) through a filter bandwidth and a windowing function of time. The second step extracts aural properties, according to the principles of ASA, required in the segmentation and grouping stages. In fact, in the segmentation and grouping, the system produces the segments for target signal and for the concurrent one then groups them from the target in a stream. This flow corresponds to a sound source from which the separated target signal will be synthesized in the last step [1, 2, 3].

CASA approach offers a rich field for the experimentation of ideas. In fact, it has numerous limitations; grouping stage is based on periodicity thus it could only be applied for voiced segments of speech. In addition, due to its dependency on pitch, the performances achieved by CASA-based approaches are affected by the accuracy of the multi-pitch estimator.

Sinusoidal Modeling (SM) of a speech signal is a spectral modeling defined as a superposition of sine waves whose frequency ratio determines the fundamental frequency of the signal [4]. It is therefore a new representation of the speech signal based on a mathematical model. To establish this model, a number of parameters that are namely the frequency, amplitude and phase is required. This may use statistical methods, pre-trained models based on Codebooks and Codevectors. The SM has shown increased efficiency in co-channel speech separation as well as single channel speech separation, not only for the harmonic signals but also for unvoiced ones. In this context, the method proposed by Macon and Clements in [5], is based on analysis-by-synthesis/overlap-add (ABS/OLA) which deals with unvoiced signals while keeping the signal characteristics. Also the approach described in [6] determines SM of unvoiced components with a minimum number of parameters to be determined and without tonal artifacts.

As SM-based methods are based on mathematical and probabilistic models, the extraction of their parameters leads to significantly complex mixture estimator algorithms and increases the complexity of the resolution. Thus they are difficult to implement in real time systems.

The decomposition into subspaces has been of great effectiveness in speech separation as well as speech enhancement and denoising [7, 8, 9].It has been first used in single channel source separation by Casey et al. [7]; the approach involves the separation of audio mixed sources based on independent subspace analysis. It proposes a method of grouping components by partitioning a matrix of independent component of cross-entropies; also known as ixegram which measures the mutual similarities of

the sources in a segment and when regrouping it, it gives the subspace sources and time trajectories. This approach has some limitations, for example the separation is carried out by finding a decomposition where the sources are statistically independent or nonredundant which is not always the case. In addition, the performance of unsupervised clustering was not sufficient hence improved in [8]. In fact, Virtanen has also developed unsupervised learning algorithm for monaural sound source separation using nonnegative matrix factorization with temporal continuity and sparseness criteria [8]. The weakness of this approach is that it requires the original signals which could not be available in most of real applications.

The prior knowledge constraint has been resolved in Molla and Hirose's approach [9]; they developed a new way of source separation from a single-mixture audio using the empirical mode decomposition (EMD) and Hilbert spectrum and without any prior knowledge about the sources [8]. But as the EMD deeply rely on derived independent basis, that are only stationary over time, good separation quality can be achieved only if the vectors are statistically independent. For example when the talkers are characterized with similar features it becomes difficult to obtain an independent basis vectors thus the separation cannot be done. In the proposed approach, the constraint of prior knowledge about the original signal is relaxed. Importantly, our approach requires only the mixture signal as input and does not make any assumption on the mixing process or about the desired speaker. In fact, we extend the Principal Component Analysis (PCA) in such a way that it achieves precise subspace separation. PCA is a powerful tool widely used in the high-dimensional data analysis as well as subspaces learning. It is based on the assumption of projecting the high-dimensional data in a linear subspace with smaller dimension. A correct estimation of the subspace allows the data size reduction and facilitates other tasks such as speech separation. The PCA process runs mainly on three essential steps; the first step substrates the mean from each of the database dimensions and calculates their covariance matrix. In the second step, eigenvectors (and eigenvalues) of the covariance matrix are extracted, thus the extracting lines that characterize the data would be possible. The last step derives the new data by choosing components and forming a feature vectors. For speech processing, subspaces decomposition could be done to construct a background model, which is represented by the mean of signal and the projection matrix comprising the first significant eigenvectors of PCA. In this way, foreground segmentation is accomplished by computing the difference between the input signal and its reconstruction.

The PCA have been extensively investigated in the field of speech processing and extended in so many ways; it has been used in [10] for speech denoising, in [11] for speech enhancement and has been added to Independent Component Analysis (ICA)

in [12] for speech separation and with Multidimensional Scaling (MDS) to determine the dimensions of speech quality [13].

Research in PCA has inspired considerable work in Robust PCA (RPCA) which combines projection pursuit ideas with robust scatter matrix estimation [14].RPCA shows a very nice framework for speech denoising but still in need of extra assumptions to achieve speech separation. Another well-known extensions of PCA is Robust Sparse PCA (RSPCA), whose robustness makes the analysis resistant to outlying observations, and Kernel PCA (KPCA) which has been used for feature extraction in speech recognition. This approach represents speech features as the projection of the mel-cepstral coefficients mapped into a feature space via a non-linear mapping onto the principal components [15]. However, it requires prior learning in order to improve classification rates.

In this paper, we extend PCA into a new way in order to separate the dominant speech signal from the intrusion or noisy speech in an unsupervised manner using sparse matrix decomposition and the low-rank technique. The proposed method directly decomposes the time-frequency matrix of composite speech into two submatrices, $X = Lo + S$, where S and Lo represent the target structure matrix and the intrusion structure matrix, respectively. This technique is able to distinguish a variable foreground from a relatively more regular background by maintaining only the real part. Then we apply a binary ideal mask. Finally, we apply a fuzzy logic classification to determine the best channel by means of our proposed multi-scale product analysis technique. The rest of this paper is organized as follows. Section 2 presents the technical details of our approach. In section 3, we report objective and subjective results. Finally, conclusions and perspective are presented in Section 4.

2. **Proposed Approach**

2.1. **Modified PCA.** The speech signal $x(t)$ whether it is a mixture of two speakers or it is a speaker altered by noise, can be described as follows:

$$x(t) = s_1(t) + s_2(t) \tag{2.1}$$

As we focus on double-talk separation problem, $s_1(t)$ and $s_2(t)$ denote the first and the second target speaker, respectively. Our goal is to extract dominant speaker from the mixture, and that is to get either $s_1(t)$ or $s_2(t)$ from $x(t)$, depending on the parameters of evaluation. Through linear transformation we can get clean speech in the transform domain. Performing the Fast Fourier Transform (FFT), we get:

$$X(n,k) = S_1(n,k) + S_2(n,k) \tag{2.2}$$

where $n = 1 : N$ and $k = 1 : K$ are the time and frequency indices. Considering the real and imaginary parts of X separately, we have:

$$Xreal(n,k) = S_1real(n,k) + S_2real(n,k) \tag{2.3}$$

$$Ximg(n,k) = S_1img(n,k) + S_2img(n,k) \tag{2.4}$$

Now we are supposed to get S_1 or S_2 from X which is a matrix decomposition problem. Considering at every moment that target speech is combined of only a limited number of frequencies, it can be modeled as a sparse matrix on spectrogram. On the other hand, compared to clean speech, mixture speech spectra within different time frames are more likely to highly correlate with each other, so we can assume intrusion to be a low-rank matrix. Now we have transformed the speech separation question into a matrix decomposition problem, which is to decompose a matrix into a sparse matrix and a low-rank matrix. In this work, we want to get the sparse and low-rank components of a matrix which is to get the solution of the following question:

$$\min_{Lo,S} rank(Lo) + \|S\|_0 \text{ s.t } X = Lo + S \tag{2.5}$$

In which $X \in \mathbb{R}^{n1*n2}$, $Lo \in \mathbb{R}^{n1*n2}$, $S \in \mathbb{R}^{n1*n2}$, $rank(Lo)$ is the rank of the matrix Lo [16], $\|\cdot\|_0$ is the Lo norm which is the number of non-zero entries in a matrix and $\lambda > 0$ is a trade-off parameter between the rank of Lo and the sparsity of S. However this is a highly non-convex optimization problem and we can obtain an optimization problem by relaxing equation (2.5) to the following convex problem:

$$\min_{Lo,S} \|Lo\|_* + \lambda\|S\|_1 \text{ s.t } X = Lo + S \tag{2.6}$$

$\|\cdot\|_*$ denotes the nuclear norm of a matrix defined as the summation of its singular values [16, 17] and $\|\cdot\|_1$ is the L_1 norm which is the sum of the absolute values of matrix entries [16]. We expect Lo to be the intrusion and S to be the clean speech and perform the decomposition as follows: First of all, we get the spectrogram of noisy speech, calculated from the FFT after overlapped framing and zero padding. Secondly, we adopt a fast and accurate algorithm for low rank and sparse decomposition, namely the inexact augmented Lagrange multiplier (IALM) technique. The recommended value of λ is as it is shown in the equation below [18, 19]:

$$\lambda = \frac{1}{\sqrt{\max(n_1, n_2)}} \tag{2.7}$$

where n_1 and n_2 are the dimensions of input matrix X. The choice of λ must adequate in order to control the amount of each subspace, and thus is related to scale issue. This issue is in fact a perennial challenge in many subspaces-decomposition problems. The IALM technique is applied to solve equation (2.6), separately given the real and imaginary parts of mixture spectrogram. Then we synthesize the two S_r and L_{or} which are the real parts of the sparse and low separated speech's spectrogram, respectively.

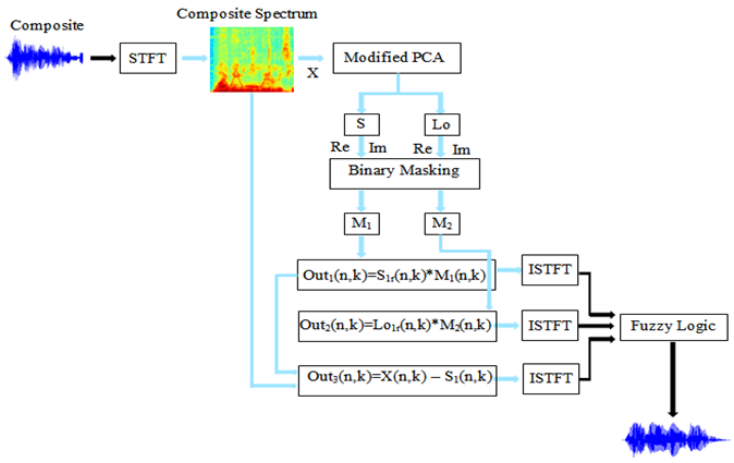

FIGURE 1. Overview of the proposed approach

In this paper, we don't apply a time-frequency masking but use the result from decomposition directly, as this can clearly reflect the statistical differences between target speech and intrusion. The diagram of the proposed modified PCA is shown in figure 2.

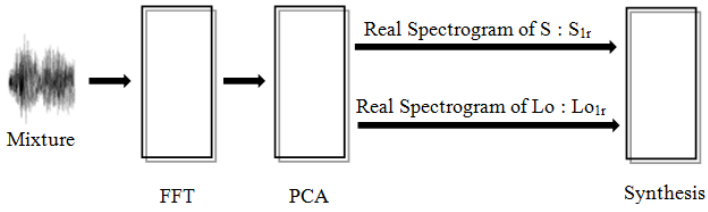

FIGURE 2. Block diagram of the proposed scheme of subspace decomposition: PCA

2.2. **Binary Masking.** Binary masking is a technique that has proven its effectiveness in speech separation as well as denoising. It allows the elimination of unwanted segments of the signal by assigning a "0" value to them, and preserves others by assigning to them a "1" value. That is therefore a hard masking (awarded label can only take 1 or 0) but there exist other masking types including the label that can take a value varying from zero to one which is called the soft masking. There exist several approaches to estimate the binary mask; those based on Bayesian classification, pitch continuity, sound localization cues and those who estimate the Posterior Signal to Noise Ratio (SNR) [20]. The binary mask $M(k, m)$ is estimated by comparing the energy of each time-frequency region of these two signals, it aims at retaining the dominant time-frequency cells in the mixture. In this work, we apply binary masking on S_{1r} and Lo_{1r} obtained from the modified PCA then we obtain two masks M_1 and M_2 and by applying them as shown in equation (8). Then the separated signals are

represented as:

$$\begin{cases} Out_1(n,k) = M_1(n,k)S_{1r}(n,k) \\ Out_2(n,k) = M_2(n,k)Lo_{1r}(n,k) \\ Out_3(n,k) = X(n,k) - Out_1(n,k) \end{cases} \tag{2.8}$$

such as Out_1, Out_2 and Out_3 are the separated signals output of the proposed approach. The speech signals will be converted back to temporal domain using the Inverse STFT (ISTFT), we obtain then three channels. The first contains out_1, the second contains out_2 and the third contains out_3, as shown in the figure 1. Choosing the desired channel will be made according to well-determined parameters by means of the Fuzzy logic, this will be discussed in the next section.

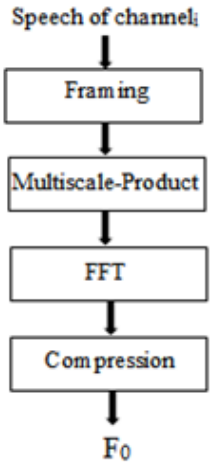

FIGURE 3. Block diagram of algorithm of fundamental frequency determining

2.3. **Candidates Selection: Fuzzy Logic.** The fuzzy nature of awarding a quality to a signal (result of a separation procedure) and the desire of simulating the behavior of the human brain have led us to opt for Fuzzy logic to do the channels classification. The first attempts of exploitation of fuzzy logic in signal processing are old but are still scarce. The fuzzy logic was used in [21] to make multilevel speech classification by developing a fuzzy voicing detector (FVD) system which allows the determining of a range between voiced and unvoiced segments. It was also used in [22] to classify audio-events in broadcast news. Despite the difference in the classification techniques using fuzzy logic in signal processing, the principle remains the same and it is based on three main steps: determining the parameters of the choice (inputs), the establishing functions and classifying rules and finally defining decisions (outputs). In our case, we have chosen the F_0 values of each channel and the average of Perceptual Quality of Speech Quality (PESQ) [23] and Signal to Noise Ratio (SNR) [23] to be the inputs.

2.3.1. *Parameters Determining.* We have applied a pitch estimation method to determine the fundamental frequency of the target speaker [24], as illustrated in figure3. After framing the input signal, the multi scale product (MP) is applied. The MP consists to compute the product of the speech wavelet transform coefficients at three successive dyadic scales. The obtained product is then weighted by a sliding window. By applying a short time Fourier transform (FFT), a peak with a clear maximum corresponding to the fundamental frequency is obtained. From the values range of F_0 we can relatively know whether the speaker is a male or is a female. We assume that if F_0 is in the range of $50 - 180Hz$ the speaker a male and if F_0 is in the range of $180 - 400Hz$ then the speaker is a female. Or in our case, the dominant speaker is a male that's why the threshold will be $180 - 400Hz$. The second parameter that influences the choice of channel is the average of SNR and PESQ, the first factor reflects the objective assessment and the second reflects the subjective assessment, then the threshold will mean the average of SNR and PESQ of mixed signal before treatment.

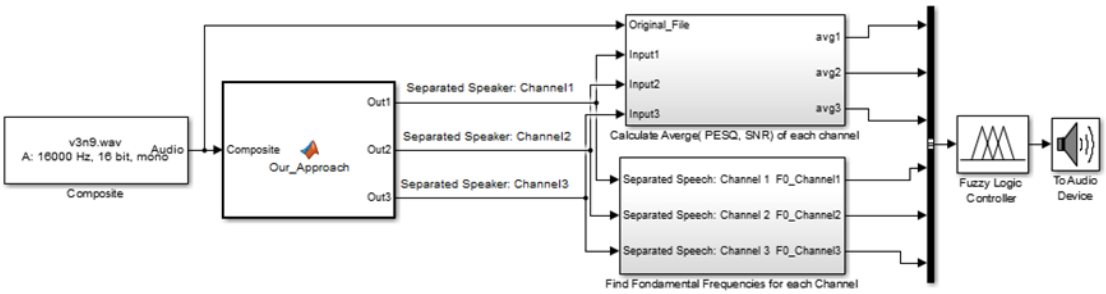

FIGURE 4. Block diagram of the Fuzzy logic scheme

2.3.2. *Classification Rules.* As the number of channels is three therefore there will be eight rules as follows:

- R1: If $F_0(\text{channel}_1) \leq F_0(\text{channel}_2)$ & $\text{Avg}(\text{channel}_1) \geq F_0(\text{channel2})$ then choose channel_1
- R2: If $F_0(\text{channel}_1) \leq F_0(\text{channel}_2)$ & $\text{Avg}(\text{channel}_1) \leq F_0(\text{channel2})$ then choose channel_2
- R3: If $F_0(\text{channel}_1) \geq F_0(\text{channel}_2)$ & $\text{Avg}(\text{channel}_1) \leq F_0(\text{channel2})$ then choose channel_2
- R4: If $F_0(\text{channel}_1) \geq F_0(\text{channel}_2)$ & $\text{Avg}(\text{channel}_1) \geq F_0(\text{channel2})$ then choose channel_1
- R5: If $F_0(\text{channel}_1) \leq F_0(\text{channel}_3)$ & $\text{Avg}(\text{channel}_1) \geq F_0(\text{channel3})$ then choose channel_1
- R6: If $F_0(\text{channel}_1) \leq F_0(\text{channel}_3)$ & $\text{Avg}(\text{channel}_1) \leq F_0$ (channel3) then choose channel_3

- R7: If $F_0(\text{channel}_1) \geq F_0$ (channel_3) & $\text{Avg}(\text{channel}_1) \leq F_0(\text{channel3})$ then choose channel_3
- R8: If $F_0(\text{channel}_1) \geq F_0(\text{channel}_3)$ & $\text{Avg}(\text{channel}_1) \geq F_0(\text{channel3})$ then choose channel_1

Such as $F_0(\text{channel}_1)$ denotes the fundamental frequencies of the channel number i and $Avg(\text{channel}_1)$ denotes the average of SNR and PESQ of the channel number i, i can take the value 1, 2 or 3. We compare the frequencies one by one. If the number of frequencies of the channel i is bigger than those corresponding to channel_j, then the channel_i is chosen. The classification strategy using fuzzy logic is given in figure 4. From which it is illustrated that the channel classification is done in three essential steps. The first step consists on getting the three outputs of the proposed approach. The second step is to compute SNR and PESQ whose average is the first parameter of the fuzzy classification. For that the original mixture is required. The second parameter is the fundamental frequency that we calculate using our MP analysis algorithm .These parameters are then fed to the fuzzy controller bloc. The strategy we have adopted for the selection of channels enabled us to make the operation automatic and even intuitive.

3. **Evaluation**

In general, the sound evaluation is performed by objective or subjective methods. On the one hand, objective methods measure the quality based on the mathematical analysis comparing the original and coded samples. The signal to noise ratio (SNR), the deformation of Itukura-Saito, the rate of logarithmic likelihood, segmental signal to noise ratio (SSNR) are among the objective methods [23]. However, to verify the accuracy of these methods, it is usually necessary to correlate with results obtained by subjective tests of speech quality. On the other hand, the subjective methods tend to measure speech intelligibility; the perceptual evaluation of speech quality (PESQ) is among these methods. In this section, we describe at first the data used in simulations to evaluate the performance of our proposed approach then the evaluation results are compared to those obtained by Hu-Wang [2], Wang-Brown [1] and Li-Guan method [3]. The evaluation will take place over two phases: an objective evaluation and a subjective evaluation. Six widely used speech quality measures were evaluated: SNR and segmental SNR (SSNR) (for objective evaluation) and perceptual evaluation of speech quality (PESQ), log likelihood ratio (LLR), weighted-slope spectral (WSS) distance (for subjective evaluation). In addition to that we used composite measures: C_{sig} for signal distortion, C_{ovl} for overall quality. C_{sig} is linear combination of LLR, PESQ and WSS, C_{ovl} is linear combination of PESQ and LLR and WSS as shown in

the equation (3.1) [23]:

$$\begin{cases} C_{sig} = 3.093 - 1.029LLR + 0.603PESQ - 0.009WSS \\ C_{ovl} = 1.594 - 0.805PESQ - 0.512LLR - 0.007WSS \end{cases} \tag{3.1}$$

3.1. **Evaluation environment.** To quantitatively evaluate our proposed approach the speech corpus used in our simulations is the Cooke database [25]; a body of composite sounds (number of 100) often used in systems analysis Computational Auditory Scene. The sounds composites are obtained by mixing ten voiced Vi speech signals, with ten interference signals Nj representing a variety of acoustic sounds, such as i vary from 0 to 9 and $j = \{1, 2, 3, 4, 7, 8, 9\}$. The signals Vi are speech signals uttered by ten male speakers. The message pronounced is the text "*Why are you all weary*". Interference is rated $N1$: white noise, $N2$: impulse noise, $N3$: Cocktail noise party, $N4$: rock music, $N7$: speech signal uttered by a woman, $N8$: speech signal uttered by a man and $N9$: signal speech delivered by a woman 2. The text of the interference corresponds to the phrase "Don't ask me to carry an oily rag like that". All signals are sampled are sampled at $16kHz$. On this basis, interference can be classified into three main categories: no interference without pitch, interference with a certain quality of pitch, speech interference.

TABLE 1. Objective Evaluation: SNR and SSNR measurements

Intrusions	N1		N2		N3		N4		N7		N8		N9	
Measures / Speakers	SNR	SSNR	SNR	SSNR	SNR	SSNR	SNR	SSNR	SNR	SSNR	SNR	SSNR	SNR	SSNR
V0	4.13	3.94	17.34	17.58	6.25	6.29	7.29	6.92	9.91	9.63	11.45	14.35	5.02	3.97
V1	2.61	3.86	14.01	12.12	3.33	5.97	7.11	6.66	9.98	7.88	13.84	12.32	4.20	4.35
V2	4.58	2.68	21.65	19.09	7.15	4.70	6.26	7.12	9.54	8.96	12.99	11.93	4.58	4.81
V3	7.55	4.79	20.77	18.88	8.28	6.86	8.26	6.62	9.16	8.43	12.32	13.99	5.58	4.80
V4	3.99	2.92	15.30	14.85	6.46	6.43	6.22	7.24	10.77	9.83	11.45	13.88	4.28	5.81
V5	3.44	3.86	14.81	15.46	6.67	8.26	4.76	8.58	8.35	8.75	12.98	11.81	4.87	3.53
V6	4.62	2.87	16.16	17.50	6.44	6.65	7.8	7.31	10.93	9.65	13.74	12.52	4.01	6.71
V7	6.82	4.98	20.47	20.39	8.86	8.92	8.68	7.52	9.38	10.55	13.57	9.94	5.90	4.28
V8	4.57	3.87	19.43	18.93	7.06	6.74	6.78	6.24	9.81	8.44	12.17	14.08	4.58	5.05
V9	4.18	3.79	14.17	16.93	6.84	6.77	6.09	7.80	8.96	7.92	12.57	11.61	4.88	7.62
Average	4.55	3.76	16.41	17.17	6.73	6.76	7.08	7.09	9.68	9.00	13.84	12.64	4.79	5.09

3.2. **Objective Evaluation.** To objectively evaluate our approach, we set two objective experiments: the signal-to-noise ratio (SNR), given in the equation (3.2), and the segmental signal-to-noise ratio (SSNR) [23].

$$SNR = \log \frac{\sum_{i=1}^{n} s(i)^2}{\sum_{i=1}^{n} (s(i) - x(i))^2} \tag{3.2}$$

such as s is the separated signal, x is the mixture and n is the length of both signals. The least square error of the separated signal and the original one is considered as

noise. The table 1 illustrates the SNR results our approach as well as the SSNR measurements. We also report the LLR and WSS measurements in table 2 and Csig and Covl measurements in table 3. The objective results show that the proposed approach that fairly good separation quality is achieved.

3.3. **Subjective Evaluation.** The performance of the approach in terms of perceived quality of the received data is evaluated by means of an objective measurement of the quality provided by the PESQ tool. This algorithm compares the received signal and the original signal and provides an objective and automated measurement for assessing the speech quality. The PESQ is characterized by the fact of being independent of the auditors and even of the number of auditors [23]. In our experiments we have used the code provided by Loizou [20]. The following table gives PESQ measurements according to our method where $PESQ_b$ and $PESQ_a$ denote respectively the PESQ values before and after treatment. Table 5 shows that our approach improves the intelligibility of the separated signals compared to original signals.

TABLE 2. Objective Evaluation: LLR and WSS measurements

Intrusions	N1		N2		N3		N4		N7		N8		N9	
Measures / Speakers	LLR	WSS	LLR	WSS	LLR	WSS	LLR	WSS	LLR	WSS	LLR	WSS	LLR	WSS
V0	3.35	242.6	1.64	83.62	1.88	52.81	3.37	62.32	2.48	47.58	1.95	107.80	1.92	152.70
V1	3.74	71.09	1.65	96.07	1.04	83.27	3.18	97.99	2.51	74.47	2.17	90.38	0.65	78.54
V2	3.70	59.80	1.54	80.18	2.07	136.4	1.46	68.54	2.93	110.04	2.23	109.26	1.96	180.51
V3	3.26	219.9	1.60	77.52	2.12	130.6	3.05	116.8	2.66	106.56	1.97	97.93	1.87	118.90
V4	3.18	237.3	1.59	76.55	1.92	125.6	3.09	150.4	2.38	99.10	1.87	87.79	1.98	125.50
V5	3.44	52.33	1.57	84.41	1.93	148.4	1.12	61.31	2.24	117.88	1.97	107.30	1.85	142.80
V6	3.32	49.46	1.71	64.59	1.78	119.2	2.89	137.20	2.19	88.28	1.59	83.95	1.89	135.10
V7	2.70	164	1.59	82.73	1.78	113.5	3.26	100.60	2.49	83.57	2.08	78.72	1.83	129.20
V8	3.03	47.86	1.55	64.65	2.18	101.4	2.74	117.50	2.19	81.83	1.64	72.86	1.62	112.50
V9	4.20	69.01	1.79	70.75	1.96	118.7	2.56	118.80	0.65	41.98	1.73	82.57	1.52	97.22
Average	3.92	121.3	1.62	78.11	1.87	113.0	2.67	103.20	2.27	85.13	1.92	91.16	1.71	127.30

4. **Discussions**

The performance of our approach has been compared to the Wang and Brown method [1], Hu and Wang method [2] and to Li and Guan method [3]. Thus we give a general overview of each one. Wang and Brown's model performs separation using the neural network. In fact it treats the ASA model from a neurocomputational perspective; it uses the oscillatory correlations as a mechanism for the ASA. This model consists of two levels: in the first level the system goes through the mixed signal and seeks an optimal source description. In the second level, the neurobiological level, recombination is done according to the features of distributed neurons. Hu and Wang approach separates voiced signals; it segregates resolved harmonics and unresolved harmonics by two different ways. For the first category, the system decomposes the

TABLE 3. Objective Evaluation: C_{sig} and C_{ovl} measurements

Intrusions	N1		N2		N3		N4		N7		N8		N9	
Measures / Speakers	C_{sig}	C_{ovl}	C_{sig}	C_{ovl}	C_{sig}	C_{ovl}	C_{sig}	C_{ovl}	C_{sig}	C_{ovl}	C_{sig}	C_{ovl}	C_{sig}	C_{ovl}
V0	2.06	0.22	2.31	2.38	0.71	2.10	1.09	1.64	0.56	2.27	1.17	2.14	5.02	1.87
V1	0.76	0.02	2.58	2.67	2.08	1.55	0.66	0.07	0.47	1.57	0.38	1.16	4.20	1.84
V2	0.58	0.19	2.49	2.51	0.61	0.74	1.77	1.42	0.31	0.14	0.86	1.09	4.58	0.37
V3	1.74	0.94	2.64	2.75	0.97	1.15	0.28	0.36	0.44	0.38	1.38	1.39	5.58	1.19
V4	1.98	1.25	2.66	2.76	1.00	1.09	0.73	0.19	0.87	1.64	1.64	1.71	4.28	1.61
V5	0.14	0.50	3.39	2.51	0.58	0.65	2.01	1.47	0.62	0.79	1.57	1.16	4.87	0.64
V6	0.06	0.49	2.64	2.78	1.13	2.10	0.40	0.18	1.07	1.28	1.96	1.92	4.01	0.97
V7	0.76	0.39	2.59	2.71	1.15	1.23	0.28	0.26	0.64	0.35	1.46	1.59	5.90	0.95
V8	0.30	0.72	2.86	2.94	0.98	1.16	0.02	0.69	1.08	1.14	1.86	1.72	4.58	1.35
V9	1.67	1.11	2.50	2.70	1.01	1.10	0.18	0.59	3.22	2.58	1.74	1.69	4.88	1.57
Average	1.01	0.54	2.67	2.67	1.02	1.29	0.74	0.69	0.93	1.21	1.40	1.66	4.79	1.24

TABLE 4. Subjective Evaluation: PESQ measurements before ($PESQ_b$) and after ($PESQ_a$) treatment

Intrusions	N1		N2		N3		N4		N7		N8		N9	
Measures / Speakers	$PESQ_b$	$PESQ_a$	$PESQ_b$	$PESQ_a$	$PESQ_b$	$PESQ_a$	$PESQ_b$	$PESQ_a$	$PESQ_b$	$PESQ_a$	$PESQ_b$	$PESQ_a$	$PESQ_b$	$PESQ_a$
V0	1.49	1.43	2.15	3.74	1.81	1.66	1.62	1.57	2.20	2.38	1.99	2.28	1.87	2.16
V1	1.00	1.79	1.48	3.29	1.32	2.99	1.30	1.47	1.67	1.65	1.91	2.13	1.47	1.60
V2	1.06	1.45	1.01	3.22	1.43	1.65	1.19	1.51	1.77	1.43	1.87	1.84	1.56	1.69
V3	1.07	1.52	1.71	3.34	1.77	1.85	1.44	1.55	2.04	2.24	1.96	2.31	1.97	1.82
V4	1.34	1.50	2.24	3.43	1.71	1.97	1.52	1.65	2.12	2.38	2.35	2.50	1.94	2.00
V5	1.38	1.39	2.07	3.87	1.25	1.67	0.84	1.38	2.05	1.89	2.07	1.94	1.83	1.56
V6	1.05	2.34	1.85	3.45	1.52	1.85	1.46	1.76	1.92	1.91	2.02	2.42	1.79	1.86
V7	0.97	2.27	1.42	3.52	1.61	1.98	1.29	1.48	1.59	1.73	1.99	2.50	1.83	2.00
V8	1.03	1.70	1.34	3.62	1.51	2.03	1.43	1.69	1.70	1.91	1.99	2.83	1.20	2.11
V9	1.41	1.84	2.25	3.42	1.74	2.28	1.67	1.97	2.22	2.51	2.35	2.41	2.25	2.3
Average	1.18	1.72	1.75	3.49	1.57	1.99	1.38	1.61	1.93	2.00	2.07	2.32	1.85	1.91

input signal into segments according to time continuity and cross-channel correlation and then groups them depending on their periodicity.However, in the second category the segmentation is done through the amplitude modulation (AM) and time continuity of and the grouping is made according to the AM rates. The separation is based on a pitch contour which is estimated according to the dominant pitch, and refined according to psychoacoustic constraints. The Li and Guan approach benefits from the CASA advantages and combines them with the objective quality assessment of speech (OQAS). The main improvement of this model compared to the basic CASA model, is the use of the OQAS in the grouping process as a guide system. The algorithm which has been chosen for the OQAS is divided into three principal stages: the preprocessing stage, the distortion estimation stage, and the perceptual mapping stage. Preprocessing stage consists on normalizing the input signal and detecting voice activity. The distortion estimation is done by tracking the pitch synchronous vocal model and linear predictive (LP) analysis, speech reconstruction and full- reference perceptual model and finally the distortion-specific parameters extraction. The third stage consists on classifying the dominant distortion and perceptual weighting. This

method is largely based on the determination of pitch. In the table 5, we compare our approach with Wang and Brown [1], Hu and Wang approach [2] and Li and Guan approach [3], in terms of SNR. We tested on all the mixed voices Vi for $\{i = 0 : 9\}$ with the intrusions Nj for $\{j = 1, 2, 3, 4, 7, 8, 9\}$ of the Cooke database. Our approach

TABLE 5. SNR comparison

Intrusions	N1	N2	N3	N4	N7	N8	N9	Average
Mixture	2.50	10.19	4.34	3.99	6.62	10.37	0.73	6.46
Proposed	**4.55**	**16.43**	**6.73**	**7.08**	**9.68**	**13.84**	**4.79**	**9.01**
Li & Guan	3.50	14.41	5.21	6.66	9.39	11.50	3.96	7.80
Wang & Brown	4.93	11.19	5.65	8.72	9.22	10.84	2.66	7.60
Hu & Wang	3.35	14.25	5.09	1.10	9.04	12.56	5.10	7.21

overcomes these methods in most of the cases. Since it does not depend on pitch, our approach can effectively separate speakers characterized margins of frequencies which comes together, even overlapping in some segments, as the male-male and the female-female mixing which are the most difficult and delicate cases of separation. Obviously, our approach is effective in the case of noise-speaker case. The stationary character of noise, which is not the case in speech, makes the decomposition into subspaces (one containing the desired speech and the other contains the noise) an easier task Figures 5 and 6 illustrate two spectrogram samples of our approach it is given respectively in the case of a noisy speech as well as the case of the two speaker's case.

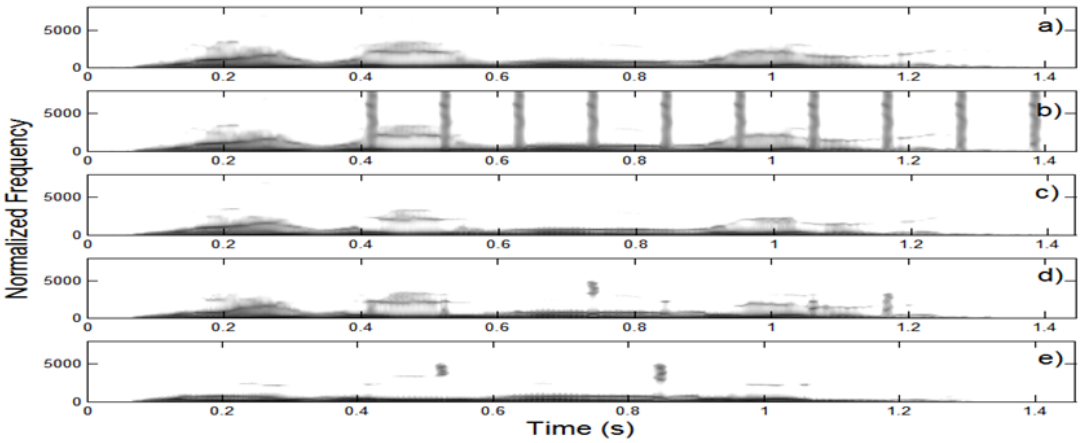

FIGURE 5. Spectrogram of enhanced speech signals in the noisy case: a) Clean b) Composite c) Proposed d) Hu-Wang e) Wang and Brown

5. **Conclusion**

In this work, we propose a speech separation approach that uses a modified PCA algorithm applied in spectral domain coupled with masking technique. This separation process results three different separated signals to be broadcast each in a channel.

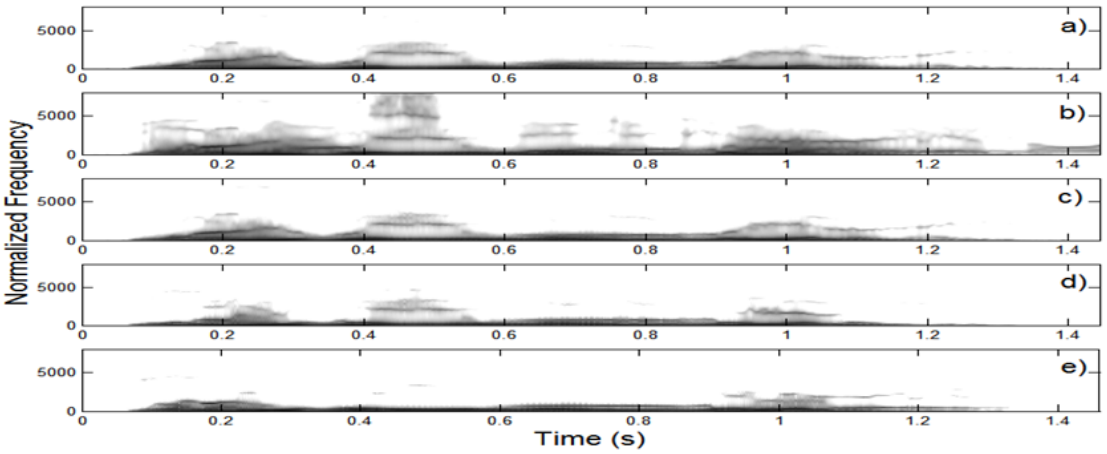

FIGURE 6. Spectrogram of enhanced speech signals in monaural speech separation case: a) Clean b) Composite c) Proposed d) Hu-Wang e) Wang and Brown

Selecting the best channel is done by means of the fuzzy logic. The contribution of our approach is that it works in a monaural environment, independently of the pitch and without any prior knowledge about the desired speaker or about the mixing process: it requires only the composite signal as input. Our approach was assessed objectively and subjectively, measures show its excellent efficiency compared to in the state-of-art approaches. Further work may address the extension of the proposed approach to optimize the decomposition of subspaces and evaluate on more than two speakers.

REFERENCES

[1] D. L. Wang, and G. J. Brown, Separation of speech from interfering sounds based on oscillatory correlation, *IEEE Trans. Neural Networks*, Vol. 10, pp. 684–69, May 1999.

[2] G. N. Hu, and D. L. Wang, Monaural speech segregation based on pitch tracking and amplitude modulation, *IEEE Trans. Neural Networks*, Vol. 15, no. 5, pp. 1135–1150, Sep 2004.

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

[4] B. Wiem, M. A. B. Messaoud, and B. Aicha, Single channel speech separation based on sinusoidal modeling, *In 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE* , pp. 672–676, Mar 2016.

[5] M. W. Macon, and M. A. Clements, Sinusoidal modeling and modification of unvoiced speech, accepted for publication *IEEE Transactions on Speech and Audio Processing*, 1997.

[6] H. Li, Y. H. Shen and J. G. Wang, Underdetermined blind separation using modified subspace-based algorithm in the timefrequency domain, *Prz. Elektrotechniczny*, Vol. 87, no. 7, pp. 280–283, Jul 2011.

[7] M. A. Casey and A. Westner, Separation of mixed audio sources by independent subspace analysis, *In Proceedings of the International Computer Music Conference*, pp. 154–161, Aug 2000.

[8] T. Virtanen, Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria, *IEEE Transactions on Speech and Audio Processing*, Vol. 15, no. 3, pp. 1066–1074, 2007.

[9] M. K. I. Molla, and K. Hirose, Single-mixture audio source separation by subspace decomposition of Hilbert spectrum, *IEEE Transactions on Speech and Audio Processing*, Vol. 15, no. 3, pp. 893–900, 2007.

[10] J. Huang, X. Zhang, Y. Zhang, X. Zou and X. Zeng , Speech Denoising via Low-Rank and Sparse Matrix Decomposition, *ETRI Journal*, Vol. 36, no. 1, pp. 167–170, 2014.

[11] T. Takiguchi and Y. Ariki , PCA-based speech enhancement for distorted speech recognition, *Journal of multimedia*, Vol. 2, no. 5, pp. 13–18, 2007.

[12] N. Kandpal and M.B. Rao , Implementation of PCA & ICA for voice ecognition and separation of speech, *In Advanced Management Science (ICAMS), IEEE*, Vol. 3, pp. 536–538, Jul 2010.

[13] D. Sen, Determining the dimensions of speech quality from PCA and MDS analysis of the diagnostic acceptability measure, *Proc. MESAQUIN*, 2001.

[14] M. Hubert, P. J. Rousseeuw and K. Vanden Branden, Determining the dimensions of speech quality from PCA and MDS analysis of the diagnostic acceptability measure, *Technometrics*, 2005.

[15] L. I. M. A. Amaro, Z. E. N. Heiga, Y. Nankaku, C. Miyajima K. Tokuda and T. Kitamura, On the use of Kernel PCA for feature extraction in speech recognition, *IEICE transactions on information and systems*, Vol. 87, no. 12, pp. 2802–2811, 2004.

[16] Y. M. A. Y. J. Wright, Sparse and Low-Rank Representation Lecture I: Motivation and Theory, *In European Conference on Computer Vision*, Oct 2004.

[17] M. Fazel, H. Hindi and S.P. Boyd, A rank minimization heuristic with application to minimum order system approximation, *In American Control Conference, IEEE*, Vol. 6, pp. 4734–4739, 2004.

[18] Z. Lin, M. Chen and Y. Ma, The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices, *UTUC technical report UILU C-ENG*, pp. 1009–5055, 2010.

[19] K. Mohamed, A. Mehdi and M. Abdelkader, Rational arnoldi & adaptive order rational arnoldi for switched linear systems, *Neural, parallel & scientific computations*, Vol. 22, no. 1-2, pp. 75–88, 2014.

[20] Y. Hu and P. C. Loizou, Techniques for estimating the ideal binary mask, *In Proc. 11th Int. Workshop Acoust. Echo Noise Control*, pp. 154–157, Sep 2008.

[21] F. Beritell, S. Casale and M. Russo, Multilevel speech classification based on fuzzy logic, *In Speech Coding for Telecommunications, 1995. Proceedings, IEEE Workshop*, pp. 97–98, Sep 1995.

[22] Z. Liu and Q. Huang and M. Russo, Classification of audio events in broadcast news, *In Multimedia Signal Processing, 1998 IEEE Second Workshop*, pp. 364–369, Dec 1998.

[23] Y. Hu and P. Loizou, Evaluation of objective measures for speech enhancement, *Proceedings of INTERSPEECH, Philadelphia, PA*, 2006.

[24] M. A. B Messaoud, A. Bouzid and N. Ellouze and M. Russo, An efficient method for fundamental frequency determination of noisy speech, *In International Conference on Nonlinear Speech Processing, Springer Berlin Heidelberg*, pp. 33–41, Jun 2013.

[25] G. J Brown and M. P. Cooke, Computational auditory scene analysis, *Comput. Speech Language*, Vol. 8, 1-2, pp. 297–336, 1994.