Received: 7 June 2023 | *Revised: 17 July 2023* | *Accepted: 27 July 2023*
DOI: 10.37917/ijeee.20.1.12 *Vol. 20 | Issue 1 | June 2024*

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Open Access
IJEEE University of Basrah College of Engineering

NNMF with Speaker Clustering in a Uniform Filter-Bank for Blind Speech Separation

Ruaa N. Ismael*, Hasan M. Kadhim
Electrical Engineering Department, College of Engineering, Mustansiriyah University, Baghdad, Iraq

Correspondance
*Ruaa N. Ismael
Electrical Engineering Department, College of Engineering
Mustansiriyah University, Baghdad, Iraq
Email: ruaa21@uomustansiriyah.edu.iq

Abstract
This study proposes a blind speech separation algorithm that employs a single-channel technique. The algorithm's input signal is a segment of a mixture of speech for two speakers. At first, filter bank analysis transforms the input from time to time-frequency domain (spectrogram). Number of sub-bands for the filter is 257. Non-Negative Matrix Factorization (NNMF) factorizes each sub-band output into 28 sub-signals. A binary mask separates each sub-signal into two groups; one group belongs to the first speaker and the other to the second speaker. The binary mask separates each sub-signal of the (257×28) 7196 sub-speech signals. That separation cannot identify the speaker. Identification of the sub-signal speaker for each sub-signal is achieved by speaker clustering algorithms. Since speaker clustering cannot process without speaker segmentation, the standard windowed-overlap frames have been used to partition the speech. The speaker clustering process fetches the extracted phase angle from the spectrogram (of the mixture speech) and merges it into the spectrogram (of the recovered speech). Filter bank synthesizes these signals to produce a full-band speech signal for each speaker. Subjective tests denote that the algorithm results are accepted. Objectively, the researchers experimented with 66 mixture chats (6 females and 6 males) to test the algorithm. The average of the SIR test is 11.1 dB, SDR is 1.7 dB, and SAR is 2.8 dB.

Keywords
Blind Speech Separation, NNMF, Filter Bank Analysis and Synthesis, Speaker Clustering.

I. Introduction

The ordinary input signal for the DSP process is a combination of information/ data signals plus additive noise signal. Information/ data signals consist of different components (e.g., the song consists of signals of the music beside the speech of the singer). For different reasons, researchers of audio, speech, and acoustic DSP processing tackle segregating these composite signals in order to recover their original signals. For instance, electronic components and recording of the audio almost add unwanted noise signals. These signals are time domain variations against the original wanted signals. Deleting or mitigating of these harmful signals is a necessary process in audio and speech DSP. This job is a segregation/ splitting method to enhance the quality of the original signals.

The segregation/ splitting method is the Source Separation process. For acoustics (speech, sound, and/ or audio) input signal, that process is speech, sound, and/ or audio separation. Due to DSP research, source separation of speech signals is more challenging than separation of audio signal. The achievements of that processing are indicative of that. Due to physical parameters (the similarity between them) and audio features of speech and speakers signals, the recovered separated speech signals have the same common characteristics and parameters between them. For the audio signals, the recovered separated signals represent the instruments, the personal sound, the machines sound etc. [1].

The researchers of audio, acoustics, and speech signal processing had enrolled with the challenge of source separation using

different methods to recover those original signals of speech (almost are defined as the targeted-speech) for each one of the speakers. According to the researchers' achievements, the well-known mathematical methods of speech separations are Independent Component Analysis (ICA) and Principal Component Analysis (PCA) matrix formulation, Computational Auditory Scene Analysis (CASA) simulation, and Non -Negative Matrix Factorization (NNMF) [2].

When the input observation signal is only one signal, it's called single-channel speech separation, which is harder than the other separation process due to the limited information about the speech and the speaker himself/ herself. Historically, the speech separation problem was named "The Cocktail-Party Problem". The speech in such a case is a mixture of more-than-one person who are talking at the same instants/ durations. If the speech separation method has more information/ attributes/ signals beside the processed observation mixture signal, it is the Informed Speech Separation. However, we do not have any of these information/ attributes/ signals, i.e., we have only the input observation speech mixture signal. In this case, the separation method is Blind Speech Separation. The current study proposes an algorithm for the blind separation of a one mixed speech signal from which the original individual signals of the two talkers' speech are extracted.

II. BACKGROUND

This paper's research exploits the abilities of the NNMF algorithms in the source separation area. To improve those abilities, the NNMF algorithm is adapted in the Time-Frequency (T-F) domain instead of the time domain directly. The speaker clustering process has been used to overcome the weak points of this adaptation.

The NNMF mathematics has excellent capability for audio separation signals. That capability is limited in the speech separation field. For source separation that was using NNMF: Hoyer made the Non-negative Spares Coding his first effort at using that method. He introduced an effective and direct multiplicative algorithm to evaluate the hidden component's optimal values. He demonstrated that it is possible to detect the observed data using the basis vectors. In [3] He also showed how explicitly incorporating the concept of sparseness improves the discovered decompositions and provided a complete MATLAB code for both standard NNMF and its extension.

By exploiting the NNMF algorithms for the audio separation, F'evotte and Ozerov used the multi-channel NNMF in convolutive mixtures [4]. In the context of their work, convolution is usually represented as an instantaneous linear combining in every frequency band of the Short-Time Fourier Transform (STFT) domain. The Itakura–Saito distance is used in the NNMF-based separation. The statistical Gaussian components represent the distance in the model. They used two methods to address estimating the source parameters and mixing. The first method is conducted via the likelihood maximizing of the multi-channel using the expectation and the maximization algorithm. The second method is executed by the likelihood maximizing of the separate model of the total channel. This algorithm is built upon the NNMF of the multiplicative method.

F'evotte, King, and Smaragdis proposed an optimization approach for NNMF-based audio, sound, and voice separation. They focused on two applications: interpolating missing musical data and single-channel source separation of speech [5]. They discussed how parameters affect performance and offered the studies' best parameters.

Kameoka, Kagami, and Yukawa introduced the "Complex NNMF (CNNMF)", an audio source separation algorithm. With this method, it is possible to build advanced time-frequency domain signal decompositions that resemble NNMF. The fact that the measure of divergence is restricted to the Euclidean distance is one of the drawbacks of traditional CNMF. In the reference, a KL divergence alternative to CNMF, which is referred to as "KL-CNMF" was presented. Moreover, a method for locating a local optimal solution was devised. They showed that KL-CNMF performed better than other traditional NNMF iterations by means of tests on supervised source separation [6].

Kadhim devised a couple of new methods: an overlapped speech detection method in addition to a couple of speech separation methods [7]. The suggested overlapped speech detection method estimates the instants of input switching. The iterations are configured to avoid poor audio features and choose the finest. The optimization is based on pattern recognition principles and k-means clustering. The suggested blind speech separation method is made up of four consecutive procedures: filter bank analysis, Non-negative Matrix Factorization, speaker clustering, and filter bank synthesis. Effective standard framing is used instead of the necessary speaker segmentation. Reasonable standard framing is added as an alternative to the necessary speaker segmentation.

Sawada, Ono, Kameoka, Kitamura, and Saruwatari explained five blind speech separation methods for audio signals. ICA and IVA rely on source independence and super-Gaussianity. NNMF and MNNMF are used to simulate spectrograms with low-rank structures. ILRMA combines these two methods and takes advantage of the independence and low rank of sources. Auxiliary function approaches can optimize all of the objective functions associated with these methods [8]. For single-channel speech separation, a Layered Convolutive NNMF (LCNNMF) technique was presented by Yao, et al. in [9]. He made a comparison between his suggested method and two

others (Non-negative Matrix Factorization (NNMF) and Layered NNMF (LNNMF)). The dataset's results demonstrated that LCNNMF performed better than NNMF and LNNMF in terms of separating the mixture of single-channel speech signals.

D. Wang, et al. proposed an NNMF-based Generalized Deep Learning Clustering (GDLC) algorithm. First, the stochastic gradient descent algorithm was used to implement the element update centered on the NNMF. The NNMF was then used to obtain the respective generalized biases and generalized weights of the two factorized matrices. The GDLC network was built by combining the activation function, the generalized biases, and generalized weights to update the respective parts of the low-dimensional matrix. The GDLC algorithm has significant advantages, according to experiments executed via eight datasets [10]. Shimada et al. presented the MNMF-guided beamforming-based unsupervised speech augmentation technique. The technique uses MNMF to calculate the unsupervised SCMs of speech and noise before generating an improved speech signal using beamforming. MNMF was made available online and initialized with an ILRMA signal and beamforming. They examined several beamforming techniques under a range of circumstances. The experimental outcomes for real-recording Automatic Speech Recognition (ASR) tasks showed that the suggested strategies were more resilient in an unknowable environment than the most advanced beamforming method using DNN-based mask estimation [11].

The NNMF system for the separation of speech sources introduced by Leplat, Gillis, and Ang in [12] is done by minimizing a cost function that contains a data fitting term (divergence) and a term for penalization that favors results [W] matrix with the lowest intensity. They demonstrated the model's identifiability in the precise scenario when the activation matrix [H] was sufficiently scattered. To address this issue, they offered multiplicative updates and demonstrated the method's behavior using audio signals from the real world. They emphasized the model's ability to handle the situation when p (factorization factor) is overvalued by automatically tuning some components to zero and producing decent source estimate outcomes.

III. The Proposed Algorithm

In this paper, the researchers only have the mixture of speech and no further database and/or information about the processed speech and the speakers. This type of source separation is called Blind Source Separation. Suppose we have n persons, in which they are talking simultaneously in a specific time period. The personal speech signal for each one of those n persons are $s_1, s_2, \ldots s_n$. These speech signals are the "targeted-speech signals" of this paper's algorithm. The speech signal (s) is the mixture of the chat between them. This conversation signal is our observation input signal for the research algorithm, which is blind speech separation [13]. Because we have only one output summation (i.e., s) of these signals for the n speakers, this speech separation is called a single channel:

$$s = \sum_{j=1}^{n} a_j s_j \tag{1}$$

ajs are scaling factors for the amplitude of each signal for those n persons. These as represent the energy content of each speech signals. For spontaneous conversations, a value is different from person to other person, from conversation to other conversations, from speech segment to other segments, from spoken sentence to other sentences, and from male to female. Generally, sometimes these variations are helpful for the speech separation process and harmful at other times. To simplify our algorithm presentation, let the number of persons enrolling with that conversation session be two. Suppose they are male m and female f speakers. This mf is the input signal of this research paper, which is the (s) signal. The researchers target to recover the separated speech signals $\hat{m}$ and $\hat{f}$ for each one alone:

$$\mathrm{m}' = \mathrm{m} + \mathrm{e_m} \tag{2}$$

$$\mathrm{f}' = \mathrm{f} + \mathrm{e_f} \tag{3}$$

e_m and e_f are the non-desired errors produced by the algorithm for the speaker m and the speaker f.

To simplify the mathematical representation of the research input signals, we have normalized the energy of the virtual signals:

$$Let\ \mathrm{a_m} = \mathrm{a_f} = 1 \tag{4}$$

$$\mathrm{mf} = \mathrm{s} = \mathrm{f} + \mathrm{m} \tag{5}$$

The above equations and formulations are the description of single-channel time-domain blind speech separation for the input observation normalized mixture speech signal.

The algorithm of this research consists of the following sequential process for the input mixture speech signal mf for the speakers m and f: filter bank (analyses the signal), NNMF

technique (factorizes the spectrogram), speaker clustering (binary mask), and then filter bank (synthesizes the sub-bands). A functional block diagram for the paper algorithm is shown in Fig. 1. At first, the input observation signal (representing mixture speech of two speakers) is processed by filter bank analysis technique to produce Short Time Fourier Transform (STFT) spectrogram for the processed speech segment. The filter band is designed for N_{sb} sub-bands output, i.e., there are sub signals N_{sb} output from the filter, each sub-band is produced by the sub-filter. By the second and the third steps, for each sub-signal, NNMF factorization produces p of sub-signals, i.e., there are N_{sb}x p sub-signals, outputs of the total factorization calculations. Using a binary mask, speaker clustering and NNMF separate the sub-signals into two speech signals groups. The first group is for the first speaker, and the other group belongs to the second speaker. The final step is synthesizing these huge amounts of sub-signal by adding the phase angle and using filter bank analysis technique.

More description of this research paper's algorithm could be presented in the following steps: At the beginning, each overlapped mixture speech segment (frame) is scaled by the standard windowing frame. The windowed-frame is overlapped with the next windowed-frame according to the references of the speech processing [14]. Windowed-Frame signal passes through filter-bank analysis of N_{sb} sub-bands. Each spectrogram of these sub-bands is the passed dynamic range of that sub-filter in the frequency domain, i.e., there areN_{sb} sub-signals, which are produced from that stage of calculations. The analysis process facilitates the mission of speech separation using the next process, which is the NNMF factorization [15]. The NNMF mathematics has excellent capability for the audio separation of signals. That capability is limited in the speech separation field. Spectral-Base of the NNMF produces N_{ss} sub-signals from each sub-band signal of the analysis stage of the filter-bank. The analysis and synthesis stages of the filter bank techniques are not sufficient to complete the entire separation process because the filter bank with the NNMF do not have the ability to identify each speaker individually. Speaker clustering process has a good alternative to complete the separation. Clustering has a binary masking effect. Phase angle should be recovered to calculate the IFFT correctly [8].

A. Filter Bank Analysis Stage

Filter bank could be defined as an array of all-pass filters in the frequency domain [14]. Basically, the filter bank was configured using analog circuits and techniques. Digital technology modified and adapted the filter. Filter bank consists of N_{sb} sub-bands of low pass, band pass and high pass filters. Low pass filter is the first sub-band filter. High pass filter is the last

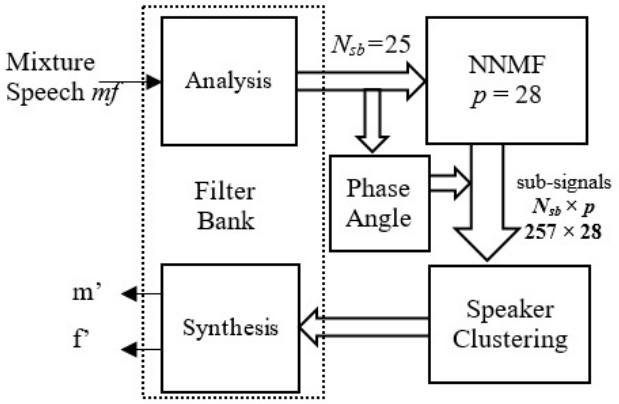

Fig. 1. The proposed algorithm: Blind speech separation.

sub-band filter (the N_{sb}^{th}). Band pass filters cover the 2nd to the (N_{sb}^{th-1}) th sub-band filters. The analysis stage of the filter bank of the research used STFT calculations, as illustrated in Fig. 2. Filter bank consists of two stages:

analysis and synthesis. Input signal of this research algorithm passes through all analysis sub filters of that filter. Specific sub-band permits the frequency domain dynamic range to pass and prevents other frequencies. Input observation signal through that analysis stage is the paper mixture speech mf. Suppose the period of mixture speech is T_{mf}. Between each frame windowed-speech (T_w) with the adjacent frames, the overlapping isT_{OL}. Hopping duration is T_{Hop}, which is the non-overlapping duration between the adjacent frames. Let the sampling rate of the speech system be fSr, which is 16k sample/ second. According to those, number of samples for total conversation is $N_{mf} = T_{mf} \times f_{sr}$; number of samples for the frame of windowed-speech is$N_w = T_w \times f_{sr}$; number of samples for the overlapping period is $N_{OL} = T_{OL} \times f_{sr}$; number of hopping period samples is $N_{Hop} = T_{Hop} \times f_{sr}$. The researchers prepared four individual segments of speech per speaker. Length of each segment is 10 second (sec) exactly (i.e., $N_{mf} = 16k \times 10 = 160k samples$). Main window length is 32 msec (i.e.,$N_w = 16k \times 32m = 512 sample$). Hopping period length is 10 msec (i.e., $N_{Hop} = 16k \times 10m = 160 sample$). Overlapping period length is 32 sec – 10 msec = 22 msec (i.e., 512 – 160 = 352 sample). Approximately, the number of frames for the conversation Nt is 160k/ 160 = 1000 frames. Due to the STFT, the number of the total sub-bands is equal to the number of input points (samples) for each main windowed frame (i.e., 512). The merits of STFT mirror conjugate in the frequency domain make it possible to reduce this number to (1 + (512 /2)) = 257 sub-band (Nsb) by:

$$[\mathrm{MF}] = \mathrm{STFT}(\mathrm{mf}) \tag{6}$$

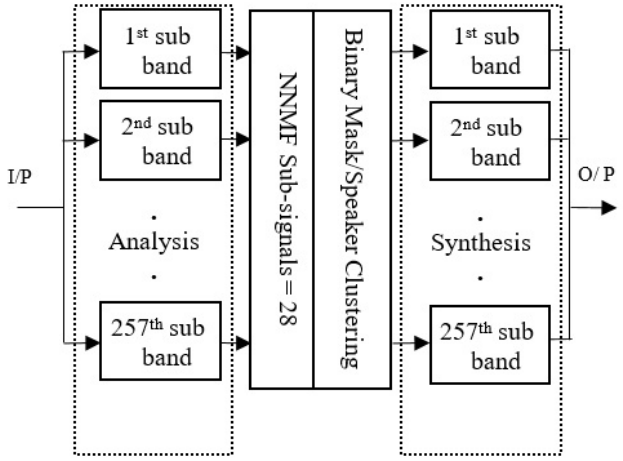

Fig. 2. The two stages of the filter bank: input I/P= Mixture speech & output O/P= Separated individual speech..

[MF] is spectrogram Time-Frequency (TF) domains matrix for our input observation mixture signal mf. There is a similarity between configurations of [MF] matrix and STFT spectrogram. For that similarity, the number of [MF] rows represents the number of sub-bands (N_{sb} = 257), and the number of [MF] columns represents the main frames (1000). Each j^{th} filter has the sub-band output:

$$[\mathrm{MF}]_{\mathrm{j}} = [\mathrm{SB}]_{\mathrm{j}} \times [\mathrm{MF}] \tag{7}$$

[MF] j is the 257×1000 STFT matrix of the jth output sub-band filter. [SB]j is the jth filter, which is a 257×257 jth sub-band square matrix. [SB]j matrix has zero elements, except the jth row elements, which are ones. Elementwise multiplication in Matlab ($.\times$) was used to manipulate the inner multiplication for the matrices. To make filter bank more efficient, the number of sub-bands is increased. This is achieved by raising the number of the main frame samples. According to speech DSP, the effective period for speech processing is (8 - 16 msec). The standard references recommended 10 msec as the better speech processing effective period. Windowing is the best compromise solution to avoid the increasing and the limitation of the duration for the main input frame of a speech signal. For this research, Hamming window scales the center of that frame to keep its energy value constant without any significant change. In contrast, the left side and the right side beside that center are attenuated severely. That scaling makes the 30 to 40 msec duration equivalent to the 8 to 16 msec effective duration. For our research, 512 STFT points sets the filter bank to 16k / 512 = 31.25 Hz per sub-band. This band width resolution is very efficient for analysis of human speech signals. To increase that efficiency, we have risen the sampling rate to 16 k samples/second, i.e., the processed speech bandwidth is 8 kHz (Nyquist-Shannon rate). Major parts of the speech components are covered by that frequency, because 4 kHz is enough bandwidth to cover deterministic human speech [14]

B. Non-Negative Matrix Factorization (NNMF)

Rectangular and square matrices are arrays, which arrange specific data and information. Specific software and hardware process this data. Day by day, these data have been expanding dramatically. That expansion needs more and more storage media expansion as well. To overcome those expansions, factorization of data matrices is feasible. Our [MF] matrix has $a \times b$ dimensions. To factorize [MF] by:

$$[\mathrm{MF}] \approx [\mathrm{W}][\mathrm{H}] \tag{8}$$

Dimensions of [W] are $a \times p$, and [H] are $p \times b$.

$$[\mathrm{er}] = [\mathrm{MF}] - [\mathrm{W}][\mathrm{H}] \tag{9}$$

[er] is an error produced from the difference between the actual values of [MF] elements and the matrix multiplication of the calculated values of [W] and [H] elements. Controlled programming iteration can be used as a factorization algorithm to calculate these values. The control is the normalization error tolerance $\|[er]\|$ or/and number of running iterations. Recently, NNMF is the best factorization algorithm to achieve this job [16]. To perform NNMF successfully, all input matrix elements [MF] are non-negative (≥ 0). The factorization matrices [H] and [W] are non-negative element matrices as well. DSP researchers have exploited the capability of the NNMF for source separation generally and for speech, sound, and audio separation particularly. For [MF] frequency domain matrix, those researchers can neglect phase angle effects. So, the —[MF]— is the absolute value of that time-to-frequency domain transformation of the speech matrix, which is the spectrogram TF-domain representation. Since all the elements of [MF] matrix are positive magnitude values, NNMF factorization could be applied to produce the two [H] and [W] matrices. The row jth in the matrix [MF] represents the vector j^{th} $[MF_j]$. The vector represents the j^{th} filter bank sub-band spectral analysis. In [MFj] vector, thei^{th} element of [MF] is:

$$\mathrm{MF_{ji}} = \sum_{\mathrm{r=1}}^{1+\mathrm{Nmf}/2} \mathrm{W_{jr}} \times \mathrm{H_{ri}} \tag{10}$$

W_{jr} is the element of j^{th} row and rth column in the matrix [W]. H_{ri} is the element of r_{th} row and j_{th} column in the matrix [H]. According to the above relationship, each frame sub-band

represents the addition result for the multiplication between the spectral base and activation weights for that sub-band. For that frame, its sub-bands spectrum is the sequential arrange of the complete sub-bands. [W] is the matrix of spectral basis for the analysis calculations. [H] is the matrix of the activation weight of those calculations. From the above calculation, the factorization factor (p) value is the number of NNMF sub-signal that were split from each sub-band signal [8]. For this research, spectral basis vector of the NNMF p equals 28. Output from NNMF is $N_{sb} \times p$ total number of sub-signals, i.e., $257 \times 28 = 7196$.

Direct substitutions and applications for the algorithm of NNMF can achieve audio separation efficiently but cannot do that for speech. The researchers proposed filter bank techniques to avoid these weak points by analyzing the main signal into Nsb sub-signals, and each sub-signal produces p sub-signals through the NNMF manipulations. The resulting number of processing/ calculations is 7196 signals. To imagine that effect, Fig. 3 arbitrarily illustrates the sub-signals produced by the filter bank analysis, which is an input to the NNNF block, and several outputs from the 28 sub-signals produced by the NNMF factorization matrices [16].

C. Speaker Clustering and Binary Masking

During the past decade, the researchers for DSP of speech had accomplished a lot of efficient speaker diarisation applications [**?**, 7]. Speaker diarisation process has two phases: the first phase is speaker segmentation, and the second is speaker clustering. Definitely, the second phase requires the first (i.e., The second cannot be processed if the first does not segment the speech signal into specific segments) [17]. For this paper, the partitioning of speech signals into standard windowed-frames is used as the speaker segmentation for that speech. The algorithm partitioned the main input conversation into the Nt main frames. Using Machine Learning (ML) labelling, the speaker clustering belongs each speech frame to its one corresponding speaker. In this research, standard speaker diarisation corpus and toolboxes are used to achieve this step successfully. According to that, NNMF factorization matrices separate the signal of any of the filter bank analysis sub-band into $p = 28$ sub signals. Each one of these 28-sub signals per sub-band is assigned to one of these two groups. The assignment for each signal is achieved but cannot be identified. Audio and speech signals can be characterized using audio features and TF tips representations. We have M and F two speakers in this research. Using traditional Euclidean distances between each one speech parameters and the reference statistical parameters, speaker clustering identifies the first group/ label to the first-person M or F. In contrast, the rest of the group should identify the second group/ label.

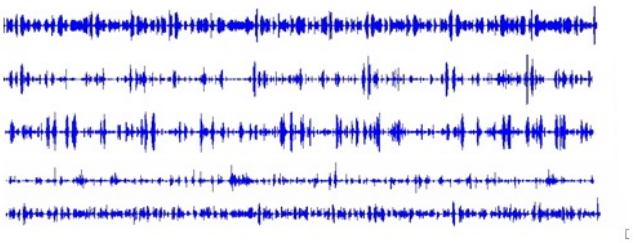

Fig. 3. 1^{st}/row is o/p from the 3^{rd} sub-band, which is i/p to NNMF. Several outputs of NNMF are the $2^{nd} - 5^{th}/rows$. The speakers are the TIMIT library [19].

Thus, the above procedure deleted (i.e., masked) the belonging of specific segment of speech out of the collection/group of one speaker, and belonged that segment to the other speaker. Such type (category) of decision is called Binary Masking, where the mask covers one label totally and uncover the other totally, i.e., share the parameters for one label and prevent these parameters from the other label [18]. Number of main windowed-frames per the input speech conversation is Nt., so the conversation has $(Nt \times 257 \times p)$ speech frames; those passes through the speaker clustering and should be shared for the first person entirely or for the second person entirely.

D. Filter Bank Synthesis

The resulting separated frames are: Nt frames per p sub-signals per 257 sub-bands. According to the masking, each frame is separated into F' (F + processing errors) or M' (M + processing errors). These separated and identified shares are the speech components of the speakers: F' and M' individually. Filter bank synthesis is used to accumulating those components. Fig. 2 shows the synthesis inputs from the NNMF algorithm, and the main output speech separated signal of one speaker. During the addition of each sub-band and each sub-signal frame with past frames results, saturation may be occurred for the vector maximum value. During the addition operation, phase-angle of the spectrograms must be retrieved from the phase-angle matrix of the mixture spectral analysis. Errors of the phase-angle is insensible for the human ears.

IV. Experiments for the Algorithm

Matlab Integrated Development Environment (IDE) was used to write the code, debug it and then implement it. The algorithm code (.m) files were written and corrected by that IDE editor. The running time for each completed .m file needed several horses to the implementation of two-persons conversation process. We have six male (M) speakers and six female (F) speakers. According to that, the prepared mixture conversations of the male-with-male MM are $(5+4+3+2+1) = 15$

mixture speech conversations. For female-with-female FF are 15 as well. For the male-with-female MF are $6 \times 6 = 36$ conversations. All the prepared speech conversations of them are $(15+15+36) = 66$. The chat is ten seconds of continuous speaking. The researchers take into consideration the different condition, sentences, energy of the virtual chat between any two persons. For that, we have chosen four different 10-second of speech per speaker. The sentences and the subject of the conversations are different from conversation to another for these four chats. There are two M and F speakers from the TIMIT standard speech library [19]. The other ten persons (five M + five F) are picked arbitrary up by the researchers from the international audio books. Most of the narrators of these books are volunteers.

To increase the resolution of filter bank, its number of sub-bands are increased by increasing the sampling rate of these selected speech to 16×103 samples/sec. To reduce quantization error of the speech, the resolution of the samples is 16-bit. Since the duration of the conversation is 10 sec, Nmf = 160,000 samples. The duration of each scaled (windowed) frame is $T_w = 32ms$, i.e., $N_w = 512$ samples $(= STFTpoints)$. Thus, number of sub-bands for the filter bank also equal 512, but the effective is $(512/2)+1 = 257$. We chose the standard time for hopping of the main frame, which is 10 msec, i.e., $N_{Hop} = 160samples$.

The number of the processed main frames are about 1000. Each sub-band of the 257 sub-bands has factorized into the $(p = 28)$ sub signals. The p has the major effect on the required time to complete the Matlab running. Increasing p factor increases number of NNMF iterations, which directly increases that running time. Increasing number of NNMF iteration decreases the NNMF [er] error, which is the priority for the researchers. For that, the researchers chose this $p = 28$ to minimize that error in spite of the very long running time.

V. Results, Tests and Comparisons

For these four trials per person, 66 conversations, the researchers implemented the algorithm according to above procedure. The four by sixty-six (244) conversations are prepared and repeated for this purpose many times to ensure that the input mixture speech chats coordinate with the terms and conditions of the standard literatures. Two of the speakers are from the TIMIT audio library [19]. Fig. 4 (B) shows the input and the output speech spectrograms, which are the Male and the Female (TIMIT) speakers. To show the differences in frequency domain, the original targeted-speech chats for them are illustrated in Fig. 4 (A) and (B) with the recovered separated output speech signals for the male and the female. After each time of the Matlab running, and for the output speech files (.wav), the subjective tests are done carefully to these conversations. Almost, the listeners denoted that major of the separated speech quality are deterministic. They noted the discontinuity of the output speech, which is caused by the binary masking. They noted the fluctuation of the separation performance. The well-knowing objective ratios (tests) are calculated for all the above output (.wav) files, which are: the SDR energy Source to Distortion test Ratio, the SAR energy Source to Artifacts test Ratio, and the SIR energy Source to Interferences test Ratio. These testes/ratios are calculated using the standard decibels (dB) meter [20, 21]. For 15 MM (Male-with-Male), 15 FF (Female-with-Female) and 36 MF (Male-with-Female) conversations, all these tests are measured to evaluate objectively these experiments. The minimum, maximum and the average values of these objective tests denoted the good capability of our research algorithm with marginal tolerance due to the different conditions of these conversations. TABLE I listed the above average values of those measurements. For that table and those average values, bar-graph figure of the objective ratios/tests are clearly shown with the details in Fig. 5. The figure illustrates the different conversations between the males, between the females, between the females-males, and for all those conversations. The Output separated speech of the female has different evaluation than the male speech outputs [19]. The comparison was between this paper's objective test with the well-known researches that used NNMF algorithms. Most of the tests for audio separation are slightly better than this research results. For speech separation, the results are in the range of these references, i.e., equal or ±10% of the averages TABLE II.

TABLE I.
Average values for SAR, SDR and SIR for the male M and female F separated speech. The MIXTURW CHATIS F with M (FM), F with F (FF), M with M (MM) and all chats.

Test	Gender	FM	FF	MM	All
SAR (dB)	F	1.87	3.15	2.77	2.69
	M	3.60	2.55	2.52	2.89
SDR (dB)	F	1.21	1.98	1.58	1.59
	M	2.03	1.70	1.47	1.73
SIR (dB)	F	12.78	10.82	10.52	11.37
	M	9.40	11.76	11.14	10.77

VI. Conclusions

According to the listeners of the subjective tests for the four times repeating of the 66 pairs of the output separated speech, the speech is deterministic. The separation is also accepted. Several of them are unaccepted and most pairs have excellent separation (the separation is very good). The three reliable SAR, SDR and SIR objective tests confirm that the algorithm

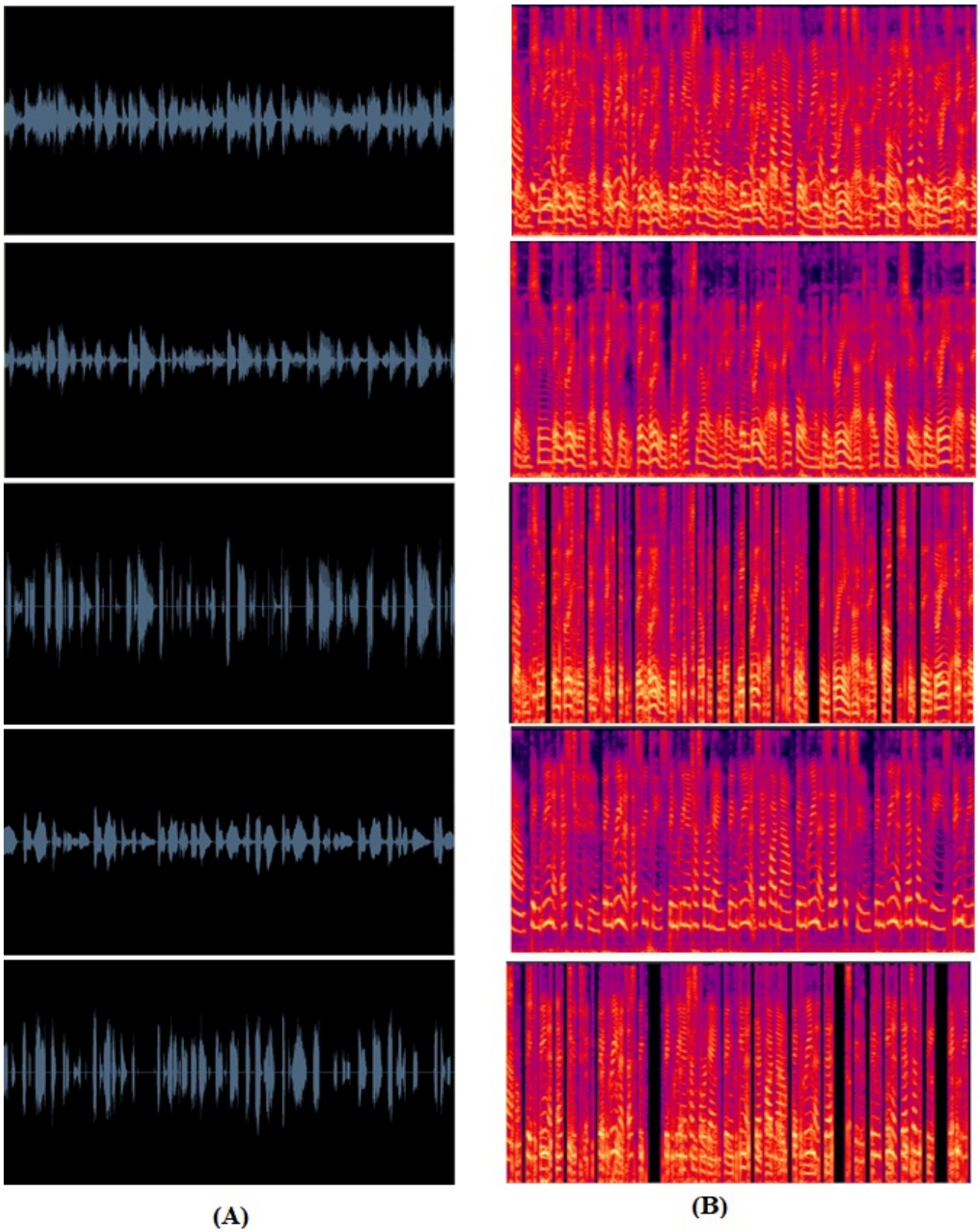

Fig. 4. Time domain TIMIT speech signal waveforms, (A) LHS column; and frequency domain spectrograms, (B) RHS column for the two TIMIT speakers, Male M and Female F. 1st/row is a mixture MF of the M with the F conversation. 2nd/row is the original M. 3rd/row is the recovered separated M'. 4th/row is the original F. 5th/row is the recovered separated F'. The discontinuity in the recovered M' and F' is due to the binary masking [19].

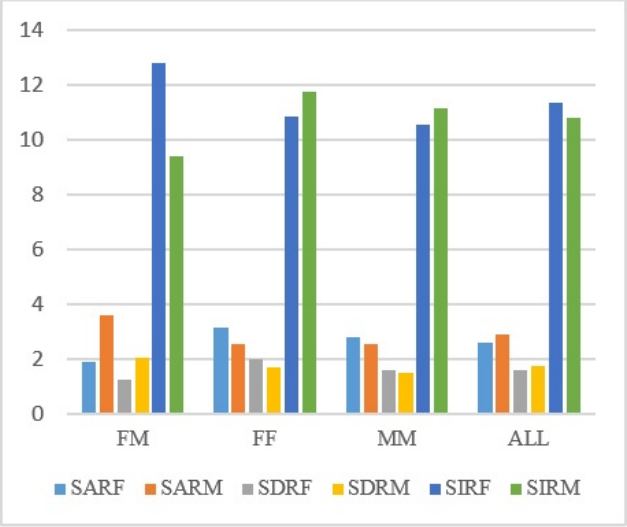

Fig. 5. Average objective tests (by dB). F for Female and M for Male (e.g., SARF is SAR for F). FM, FF, MM and ALL are the mixture (FM is F with M mixture, FF is F with F mixture, MM is M with M mixture, and ALL is for all these chats.

TABLE II.
THAN OUR RESEARCH, BUT ±10% FOR THE SPEECH SEPARATION.

SDR	SAR	SIR	Reference
1.0-4.0	4.5-8.5	-	[5]
0.1-3.9	(-0.9)-5.2	4.0-8.0	[22]
0.8	3.3	10.96	[23]
2.89	4.59	12.72	[23]
0.51-2.8	-	-	[24]
2.2-3.5	-	-	[25]
1.38	1.38	1.9	[26]
4.02	-	-	[27]
3.3 - 4.6	5.8 - 13.2	(-1.3) - 6.8	[28]
1.5 - 6.3	-	1.5 - 6.3	[29]
1.14 - 9.60	7.78 - 10.09	0.12 - 19.8	[12]
4.69 - 7.89	2.33 - 14.98	(-1.73) - 8.9	[12]
(-4.2) - 7.9	2.64 - 15.2	(-1.39) - 9.0	[12]
1.70	2.80	11.1	This article

has good efficiency for the blind speech separation. Compared with other recent articles, the algorithm is efficient. The weak point of the algorithm is the time required for running the Matlab to perform the calculations. To reduce that time, number of sub-bands of the filter bank and/or number of sub-signals of the NNMF could be reduced. 257 sub-bands are enough to define the resolution of frequency domain properly. 20-30 sub-signals is enough for the NNMF factorization of the input sub-bands. Sub-band × sub-signal must be several southlands to achieve the separation efficiently. More than that, increase the efficiency but expands the required time rapidly. Accordingly, the entire algorithm process is implemented, so the duration of the mixture of conversation segment should be reasonable, i.e., at least several seconds. The algorithm is done for 2-speaker conversations, so it should be repeated for 3 or more speakers. The consideration of the conversation subjects is important. The main factor for the failure and the successful of the algorithm is the number of filter bank sub-bands and NNMF sub-signals. The flexibility of increasing and/or decreasing should be investigated carefully.

ACKNOWLEDGMENT

We would like to express our gratitude and respect to our Department of Electrical Engineering. Appreciation also goes to our University of Mustansiriyah.

CONFLICT OF INTEREST

The authors have no conflict of interest relevant to this research.

REFERENCES

[1] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[2] D. Raj, P. Denisov, Z. Chen, H. Erdogan, Z. Huang, M. He, S. Watanabe, J. Du, T. Yoshioka, Y. Luo, *et al.*, "Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis," in *2021 IEEE spoken language technology workshop (SLT)*, pp. 897–904, IEEE, 2021.

[3] P. O. Hoyer, "Non-negative sparse coding," in *Proceedings of the 12th IEEE workshop on neural networks for signal processing*, pp. 557–565, IEEE, 2002.

[4] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio

source separation," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 3, pp. 550–563, 2009.

[5] B. King, C. Févotte, and P. Smaragdis, "Optimal cost function and magnitude power for nmf-based speech separation and music interpolation," in *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, IEEE, 2012.

[6] H. Kameoka, H. Kagami, and M. Yukawa, "Complex nmf with the generalized kullback-leibler divergence," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 56–60, IEEE, 2017.

[7] H. M.-A. Kadhim, *Single channel overlapped-speech detection and separation of spontaneous conversations*. PhD thesis, 2018.

[8] H. Sawada, N. Ono, H. Kameoka, D. Kitamura, and H. Saruwatari, "A review of blind source separation methods: two converging routes to ilrma originating from ica and nmf," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e12, 2019.

[9] W. Yao, D. Lv, X. Huang, J. Zi, M. Gao, R. Xi, and Y. Zhang, "Layered convolutive nonnegative matrix factorization for speech separation," in *Journal of Physics: Conference Series*, vol. 2258, p. 012020, IOP Publishing, 2022.

[10] D. Wang, T. Li, P. Deng, F. Zhang, W. Huang, P. Zhang, and J. Liu, "A generalized deep learning clustering algorithm based on non-negative matrix factorization," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 7, pp. 1–20, 2023.

[11] K. Shimada, Y. Bando, M. Mimura, K. Itoyama, K. Yoshii, and T. Kawahara, "Unsupervised speech enhancement based on multichannel nmf-informed beamforming for noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 5, pp. 960–971, 2019.

[12] V. Leplat, N. Gillis, and A. M. Ang, "Blind audio source separation with minimum-volume beta-divergence nmf," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3400–3410, 2020.

[13] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, IEEE, 2021.

[14] L. Rabiner and R. Schafer, *Theory and applications of digital speech processing*. Prentice Hall Press, 2010.

[15] C. Févotte, E. Vincent, and A. Ozerov, "Single-channel audio source separation with nmf: divergences, constraints and algorithms," *Audio Source Separation*, pp. 1–24, 2018.

[16] H. A. Song and S.-Y. Lee, "Hierarchical representation using nmf," in *Neural Information Processing: 20th International Conference, ICONIP 2013, Daegu, Korea, November 3-7, 2013. Proceedings, Part I 20*, pp. 466–473, Springer, 2013.

[17] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal processing*, vol. 88, no. 5, pp. 1091–1124, 2008.

[18] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. Le Roux, and J. R. Hershey, "Universal sound separation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 175–179, IEEE, 2019.

[19] S. Fernández, A. Graves, and J. Schmidhuber, "Phoneme recognition in timit with blstm-ctc," *arXiv preprint arXiv:0804.3269*, 2008.

[20] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, V. Gowreesunker, D. Lutter, and N. Q. Duong, "The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928–1936, 2012.

[21] P. Mowlaee, R. Saeidi, M. G. Christensen, and R. Martin, "Subjective and objective quality assessment of single-channel speech separation algorithms," in *2012 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pp. 69–72, IEEE, 2012.

[22] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1825–1828, IEEE, 2008.

[23] R. Jaiswal, "Non-negative matrix factorization based algorithms to cluster frequency basis functions for monaural sound source separation.," 2013.

[24] B. Gao, W. L. Woo, and S. S. Dlay, "Unsupervised single-channel separation of nonstationary signals using gammatone filterbank and itakura–saito nonnegative matrix two-dimensional factorizations," *IEEE Transactions*

on Circuits and Systems I: Regular Papers, vol. 60, no. 3, pp. 662–675, 2012.

[25] B. Gao, W. L. Woo, and S. S. Dlay, "Variational regularized 2-d nonnegative matrix factorization," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 5, pp. 703–716, 2012.

[26] B. Gao, W. L. Woo, and S. S. Dlay, "Adaptive sparsity non-negative matrix factorization for single-channel source separation," *IEEE journal of selected topics in signal processing*, vol. 5, no. 5, pp. 989–1001, 2011.

[27] A. Al-Tmeme, W. L. Woo, S. S. Dlay, and B. Gao, "Underdetermined convolutive source separation using gemmu with variational approximated optimum model order nmf2d," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 35–49, 2016.

[28] G. Cantisani, S. Essid, and G. Richard, "Neuro-steered music source separation with eeg-based auditory attention decoding and contrastive-nmf," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 36–40, IEEE, 2021.

[29] A. Alghamdi, G. Healy, and H. Abdelhafez, "Real time blind audio source separation based on machine learning algorithms," in *2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pp. 35–40, IEEE, 2020.