IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 June 2022, accepted 26 June 2022, date of publication 5 July 2022, date of current version 30 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3188712

RESEARCH ARTICLE

Speech Separation Using Augmented-Discrimination Learning on Squash-Norm Embedding Vector and Node Encoder

HA MINH TAN[1], KAI-WEN LIANG[1], YUAN-SHAN LEE[1], (Member, IEEE), CHUNG-TING LI[2], YUNG-HUI LI[2], (Member, IEEE), AND JIA-CHING WANG[1], (Senior Member, IEEE)

[1]Department of Computer Science and Information Engineering, National Central University, Taoyuan 320317, Taiwan
[2]AI Research Center, Hon Hai Research Institute, Taipei 114699, Taiwan

Corresponding author: Jia-Ching Wang (jcw@csie.ncu.edu.tw)

This research is partially supported by MOST 110-2634-F-008-004, MOST 110-2634-F-008-008, MOST 111-2321-B-195-001, MOST 110-2634-F-006-022, and MOST 110-2221-E-008-076-MY3.

ABSTRACT Speech separation has been employed in important applications such as automatic speech, paralinguistics, speech recognition, hearing aids, and human-machine interactions. In recent years, deep neural networks have been widely used for speech and music separation. Some of these breakthrough successful models based on embedding vectors have been proposed, such as deep clustering. In this paper, we propose a node encoder Squash-norm deep clustering (ESDC) as an enhanced discriminative learning framework by combining node encoder, Squash-norm, and deep clustering (DC). First, a node encoder is used to create intermediate features. Node encoders are developed through a matrix factorization-based learning method for graph representations. It creates distinguishable intermediate features that play an important role in improving performance. These discriminated intermediate features are then used as input features for the separation block. The decoder block finally constructs the estimation mask through the clustering method and reconstructs the estimated signal for each source. In particular, we apply a normalization function, Squash-norm, to the input and output vectors to enhance the distinction between high-dimensional embedding vectors. This nonlinear function amplifies the differences in the input vectors, resulting in highly unique features, which are scalar products of the vectors. Similar to the input vector, Squash-norm also enhances the discrimination of the output vector, thereby enhancing the ability to construct an estimated mask by clustering the output vector. Overall, the proposed ESDC achieves 1.27–2.09 dB SDR, 1.28–2.21 dB SDRi, and 1.3–2.44 dB SI-SNRi gain compared to the DC baseline separation performance across genders on the TSP and TIMIT datasets. With the same gender, our proposed ESDC achieves 1.14–2.71 dB SDR, 0.99–2.74 dB SDRi, and 0.62–2.86 dB SI-SNRi gain compared with the DC baseline on the TIMIT dataset. In all cases, the proposed ESDC model consistently maintains STOI and PESQ higher than the DC baselines on the TSP and TIMIT datasets.

INDEX TERMS Speaker separation, supervised speech separation, monophonic source separation, speech enhancement, time frequency masking, deep clustering.

I. INTRODUCTION

Audio signals provide a vast amount of important information by which human hearing can easily distinguish specific speech sources in a multi-speaker environment, such as at a cocktail party where multiple voice sources are blended together. However, source separation on a computer is a challenging task, especially in single-channel source separation, which is an extremely difficult task that uses only one single microphone to collect source signals.

In recent decades, statistical models, probabilistic models, clustering methods, and factorization methods have been

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

widely used for acoustic decomposition and many applications. Clustering techniques include probabilistic distributions such as the Gaussian Mixture Models (GMMs) [1], and the hidden Markov (HMMs) [2], as well as statistical models, e.g., the independent component analysis (ICA) [3]. Traditional methods such as computational auditory scene analysis (CASA) [4], [5], developed from cognitive psychology and spectral clustering [6], separate points into discrete clusters based on the eigenstructure of affinity matrices. However, the statistical models, probabilistic models, and clustering methods all lack generalization and have restrictions on the observed elements [7]. Factorization techniques, such as the Nonnegative matrix factorization (NMF) [8], [9], are regarded as an outstanding source decomposition method, however, NMF is only practical for a small number of facilities and real-time applications are difficult to realize due to the complex isolation processes of a large number of bases.

In recent years, deep learning models have been widely used for computer vision [10], [11]. Deep learning methods have been used for source separation lately, such as deep neural networks (DNNs), recurrent neural networks (RNNs) [12], [13], and convolutional networks (CNNs) [14]. In particular, DC [15] is proposed to attain robust source separation performance in the time-frequency (TF) domain, and specifically addresses the problem caused by permutation. During the training phase, DC is used to map the spectrum of a mixture to the embedding matrix. During the test phase, individual speakers in the mixed signal are isolated by clusters implemented using a binary mask of an embedding matrix. Several updated versions of DC have improved performances [16]–[21]. Related works have been successively proposed to improve the performance. The deep attractor network (DANet) [22] uses attractor points in which the weights of the DANet are learned from the deep embedding space. The permutation invariant training (PIT) [23] uses the mean square error (MSE) method to solve the problem resulted from source permutation. An utterance-level PIT (uPIT) [24] further improved PIT by fixing each talker to an output layer throughout each training utterance. Causal deep CASA [25] achieves state-of-the-art separation performance with fixed and arbitrary number of sources. A single-channel speech dereverberation algorithm [26] is proposed to resist the effect of reverberation, it uses a temporal convolutional network (TCN) architecture. The causal voice separation approach [27] has the simultaneous occurrence of reverberation, which is the fundamental requirement for real-time operation. The models perform vocal separation in the time domain directly and obtain state-of-the-art results such as the time-domain audio separation networks (TasNet) [28], Wavesplit [29], FurcaNet [30], Wave-U-Net [31], LaFurcaNet [32], the dual-path RNN (DPRNN) [33], and the gated DPRNN [34]. The architectures of intra-segment and inter-segment of DPRNN are the effective approaches that are applied to the recent effective methods, e.g., DPTNet [35] and SepFormer [36]. In addition, lightweight approaches based on DPRNN have been introduced, e.g., a self-attentive network with a novel sandglass-shape (Sandglasset) [37], selective mutual learning approach (SML) [38], GroupComm-DPRNN [39], etc.

In this paper, we propose the ESDC model, which improves the discriminative ability of high-dimensional vectors for monophonic speech separation. ESDC achieves impressive separation performance in the TF domain. The ESDC performs four stages, such as input feature encoding, embedding vector training, vector normalization, and vector clustering. In the input feature encoding stage, the node encoder transforms the input feature vectors into scalar product features, thereby creating correlation of neighboring information. The node encoder block uses the adjacency-based similarity of the embedded feature matrix from the input feature vector to establish the relationship between features. The scalar product features are proportional to the strength of the relationship between the input vectors. The high discriminative rate of scalar product features is an important feature to improve the performance of source separation in the training stage. During the embedding vector training stage, we use the backbone network to train scalar product features. In the vector normalization stage, Squash-norm is used to enhance the discriminative ability of deep-dimensional feature vectors. When the vector is short, the vector becomes close to the zero vector. Conversely, if it is a long vector, the vector becomes close to a unit vector. In the vector clustering stage, various clustering algorithms are used to cluster the embedding vectors. The main contributions of our proposed ESDC method are summarized as follows:

1) We propose a discriminative vector learning strategy, namely ESDC, for single-channel speech separation. It can enhance the discriminative learning ability of the embedding vectors.
2) We use the node encoder to generate the scalar product features, the backbone network trains the embedding vector, and Squash-norm enhances the discriminative ability of the embedding vectors.
3) The experimental results show that our proposed ESDC method achieves state-of-the-art separation performance for the same gender and different gender in the T-F domain. This approach has the ability to separate voices with known or unknown speaker numbers. The details are discussed in Section VI.

The rest of the paper is organized as follows. The related work is presented in Section II. Section III presents the problem formulation of monophonic speech separation. Section IV describes the mask and training criteria. Section V presents the proposed ESDC model. Section VI discusses the experimental results and Section VII concludes this paper.

II. RELATED WORK

Single-channel speech separation isolates distinct sources from a single recording with overlapping source sounds. This is a classic work in signal processing [41], [42], which has developed rapidly in recent years thanks to supervised neural networks. Traditionally, TF mask-based learning has been

TABLE 1. Summary of advantages and disadvantages of the models.

Models	Advantages	Disadvantages
DC [15]	– A novel approach trains embedding vectors during the training phase and clusters these vectors in the testing phase. – The loss function is the difference between the embedding affinity matrix and binary affinity matrix. It addresses permutation invariant training. – Good results were obtained with both fixed and arbitrary source cases.	– The size of the embedding dimension increases lead to an increase in computational complexity. – Loss function solves the permutation problem well, but makes it computationally complex. – The reconstruction lacks phase information for each TF bin, leading to errors and artifacts in the estimated source.
DC++ [17]	– DC++ extends DC and proposes some improvements to make the algorithm better. – The STFT performs global mean-variance normalization as a preprocessing step before being used as an input feature. – The model incorporates a dropout to avoid overfitting and to allows for a higher initial learning rate. – The soft k-means method is used to obtain a more flexible TF bin assignments. – The last improvement is to use an end-to-end approach instead of the clustering stage. This result is better performance when the number of the sources is known.	– Some techniques are added to enhance performance, however these techniques have been discovered before. – This model reuses the DC architecture, and only changing the number of layers and the number of nodes per layer.
DC with gated convolutional network [20]	– CNN architectures are used instead of BLSTM layers. – Dilated convolution outperforms the separation performance of BLSTM layers and other traditional CNNs. Dilated convolutional models provide the same results even with smaller training datasets. The backbone network with dilated convolution layers reduces computational cost .	– Performance is not significantly improved. – DC with gated convolutional network is mainly inherited from DC. – Several CNN architectures have been investigated, but novel architectures have not been proposed.
Single hybrid network [40]	– DC and traditional network architectures seem to complement each other. The authors refer as Chimera network. – The hybrid network, which combines a deep clustering network and a traditional network, is trained using a multi-task learning approach. This hybrid model significantly outperforms any of its components.	– The Chimera network almost inherits DC++ and traditional networks. – The training time and computational cost increase for the single hybrid network . – Phase information for each TF bin isn't used to reconstruct the estimated sources.
Improved hybrid network [16]	– An improved hybrid model combines DC and mask-inference networks using multiple alternative loss functions. – Alternative loss functions include Graph Laplacian distance loss, deep clustering loss with Stochastic normalization, deep linear discriminant analysis (Deep LDA) loss, deep clustering loss with whitened k-means, and deep clustering loss with introducing weights. – Phase information is adopted for time-domain re-synthesis.	– Improved hybrid network is mainly inherited from Chemira network [40]. – Several novel loss functions are proposed. The complexity of the loss function leads to better performance but increases in computational cost.

used in these models. The input mixture contains TF bins and each TF bin of the source with the highest energy is found. Spectrograms for each source can be created by masking the TF bins of the other sources.

DC [15] approach provides a clustering model for masking. The model trains the embedding vectors such that the distance between the embedding vectors from the same source is smaller than the distance between the embedding vectors from different sources. In [17], the authors propose global mean-variance normalization as a preprocessing step on the STFT spectrogram, which is then used as input to the model and the estimated mask is created by soft k-means. In [20], the authors use gated convolutional layers instead of the BLSTM layers and the dense layer. A model adopts a new loss function, while the structure of the model is preserved in [16]. A network structure with two heads [40], this hybrid network is designed to incorporate both DC and the inference mask network. The body of the network is trained using a combination of the losses from the two heads. These methods train embedding vectors as the DC method and improve performance significantly. In this paper, we propose the ESDC method, that outperforms the deep clustering method. In the ESDC architecture, the node encoder is used to create scalar product features, the backbone network is used to train the embedding vector, and Squash-norm is used to improve the discriminative ability of the embedding vector.

III. PROBLEM FORMULATION OF MONOPHONIC SPEECH SEPARATION

The goal of source decomposition is to reconstruct the single-channel source signals from the mixture signal. The discrete-time mixed signal $x(n)$ of C single-channel source signals is denoted as $x(n) = \sum_{c=1}^{C} s_c(n)$, $s_c(n) \in \mathbb{R}$. In this paper, the short-time Fourier transform (STFT) with the constant overlap-add [43]–[45] is used to transform the original signal. The STFT divides a long-time signal into the short frame signal and then computes the Fourier transform on each short frame, and can be written as follows:

$$X(t,f) = \sum_{n=0}^{L-1} x(n)\, w(n - Ht)\, e^{-j2n\pi f/L} \in \mathbb{C}, \quad (1)$$

where $w(n)$ represents the window function of L-points discrete Fourier transform (DFT) with H frame shift, $X(t,f)$ represents the STFT of the mixed signal $x(n)$, and $t, f \in \mathbb{N}$, $1 \leqslant t < T$, $1 \leqslant f < F$ are the time frame, and the frequency bin in the STFT, respectively. Due to the characteristics of STFT, a mixed signal $X(t,f)$ in the TF domain is calculated from the single source signals by

$$X(t,f) = \sum\nolimits_{c=1}^{C} S_c(t,f) \in \mathbb{C}, \tag{2}$$

where $S_c(t,f)$ represents the STFT coefficient of the component source signal $s_c(n)$ in the mixed signal. The signal is represented by the amplitude and phase $S_c(t,f) = |S_c(t,f)| \exp(j\theta_c)$. The phase of the STFT is usually ignored [9], [46] so the spectral magnitude $|X(t,f)|$ of the mixed signal can be approximated by the sum of the spectral magnitudes of the single sources as follows [47], [48]:

$$|X(t,f)| \approx \sum\nolimits_{c=1}^{C} |S_c(t,f)|, \tag{3}$$

where $|S_c(t,f)|$ is the spectral magnitude of the single source. The target mask, $M_c(t,f) \in \mathbb{R}$, is used as the target of the deep neural network in the magnitude domain. The estimated signal, $\widehat{S}_c(t,f)$, of each source signal can be calculated by

$$\widehat{S}_c(t,f) = \widehat{M}_c(t,f) \odot X(t,f), \tag{4}$$

where $\odot$ represents the element-wise multiplication and $\widehat{M}_c(t,f)$ represents the estimated mask. The estimated signal $\widehat{S}_c(t,f)$ in the TF domain is used to reconstruct each estimated source signal $\widehat{s}_c(n)$ in the time domain by inverse short-time Fourier transform (iSTFT), and can be calculated as follows:

$$\widehat{s}_c(n) = \frac{1}{L\omega(n-Ht)} \sum_{f=0}^{L-1} \widehat{S}_c(t,f) e^{j2n\pi f/L} \in \mathbb{R}. \tag{5}$$

IV. MASK AND TRAINING CRITERIA

A. IDEAL BINARY MASK

The ideal binary mask (IBM) in [49] is a TF mask constructed from the component source signal in the mixed signal. The IBM is utilized as a computational target for CASA and the training criteria of neural networks for source separation in the TF domain. For each TF node, if the local $SNR(t,f)$ is greater than the local criterion (LC), the mask value is one. Conversely, if the local $SNR(t,f)$ is less than the LC, it is zero. The IBM is defined as:

$$IBM_c(t,f) = \begin{cases} 1, & \textit{if } SNR(t,f) > LC \\ 0, & \textit{otherwise}, \end{cases} \tag{6}$$

where $SNR(t,f)$ is the local ratio of the two source signals in the mixture within the TF node.

B. IDEAL RATIO MASK

The ideal ratio mask, (IRM), is shown as a soft version of the IBM and is defined as follows [49], [50]:

$$IRM_c(t,f) = \frac{|S_c(t,f)|}{\sum\limits_{c=1}^{C} |S_c(t,f)|}, \tag{7}$$

where $|S_c(t,f)|$ is the spectral magnitude of the single source. The IRM is limited to $0 \leqslant IRM_c(t,f) \leqslant 1$ and $\sum_{c=1}^{C} IRM_c(t,f) = 1$ for all TF nodes. The IRM is similar to the classic Wiener filter and regarded as the optimal estimation tool of target speech from the standpoint of power spectrum.

TABLE 2. Math symbols in problem formulation of single-channel speech separation.

Symbol	Meaning/definition
t	Time frame in the short-time Fourier transform
f	Frequency bin in the short-time Fourier transform
T	Number of time frames
F	Number of frequency bins
C	Number of source signals
D	Deep dimension value
$w(n)$	Window function of discrete Fourier transform
$x(n)$	Discrete-time mixture signal
$X(t,f)$	Mixture signal in the time-frequency domain
$s_c(n)$	Time-domain source signal
$S_c(t,f)$	Source signal in the time-frequency domain
$\widehat{S}_c(t,f)$	Estimated source signal in the time-frequency domain
$\widehat{s}_c(n)$	Estimated source signal in the time domain
$M_c(t,f)$	Target mask in the time-frequency domain
$\widehat{M}_c(t,f)$	Estimated mask in the time-frequency domain
$\|X(t,f)\|$	Spectral magnitude of the mixed signal
$\|S_c(t,f)\|$	Spectral magnitude of the single source
$IBM_c(t,f)$	Ideal binary mask
$IRM_c(t,f)$	Ideal ratio mask
$PSM_c(t,f)$	Phase-sensitive mask
$LN(X)$	Layer normalization of the mixed signal
$\mu(X)$	Mean of the mixed signal
ϕ	Difference phase
$\sigma(X)$	Standard deviation of the mixed signal
ε	Small positive number
z_i, z_j	Input vectors
v_i, v_j	Output vectors
u_i, u_j	Output vectors with Squash-norm
y_i, y_j	Indicator vectors
$\ddot{u}_i, \ddot{u}_j$	Output vectors with unit-norm
m_c	Centroid of the c^{th} cluster
τ_{ic}	Element of a label vector τ_i
e_{artif}	Artifacts
e_{noise}	Noise
e_{interf}	Interferences
d_{SYM}	Symmetric disturbance
d_{ASYM}	Asymmetric disturbance
Z	Input embedding matrix
S	Scalar product of input vectors
Y	Label mask
V	Output matrix
U	Output normalization matrix

C. PHASE-SENSITIVE MASK

The phase-sensitive mask (PSM) reflects the relationship between the clean voice, the mixed voice, and the phase difference, which is given by

$$PSM_c(t,f) = \frac{|S_c(t,f)|}{|X(t,f)|} \cos\phi, \tag{8}$$

where ϕ denotes the difference between the pure voice phase and the mixed-voice phase in the TF bin. The PSM is extended from the IRM and provides extra phase information.

D. TRAINING CRITERIA

Since we use the scalar product of indicator vectors as the training criteria in the proposed ESDC model (to be defined in detail in Sec. V). The TF mask, $M_c(t, f)$, is utilized to construct the indicator vectors, $y_i, y_j \in \mathbb{R}^{1\times C}$. Let $y_{i,c} \in \mathbb{R}^{1\times 1}$ denote that each element of the indicator vector corresponds to the TF unit of the reference mask. In the experiments, $IBM_c(t, f)$, $IRM_c(t, f)$ and $PSM_c(t, f)$ are used to construct each element of the indicator vector respectively, in the cases.

V. THE PROPOSED MODEL FOR MONOPHONIC SPEECH SEPARATION

Although DC achieves good performance in the source separation tasks in the TF domain, it is still worth looking for further improvements. In this paper, we propose the ESDC model is shown in Fig. 1. We have two main contributions to the separation performance. First, the node encoder alters the mixture spectra of the input vectors, resulting in the highly distinctive features which are the scalar product of the input vectors. These features are helpful to improve the performance in the training stage. Second, Squash-norm ensures discriminative learning among the embedded vectors by shrinking the vectors with small norm and dilating the large ones. The objective function optimizes the feature representation by amplifying the difference between the scalar products of the normalized deep-dimensional output vectors of the model and the scalar products of the indicator vectors.

A. PREPROCESSING BLOCK

In preprocessing block, we use STFT with constant overlap-add (Eq. 1) to transform the discrete-time mixture speech sequences and detailed in Sec. VI-A1. We perform logarithm on the magnitude of the STFT coefficients. The transformed log-scaled magnitude spectra are then normalized in two standard score layers, a global mean-variance layer for all spectra of the dataset and the local mean-variance layer for the magnitude spectra, $X_i(t, f)$, of each segment in the dataset. The general formula used to these layers is given by

$$LN(X) = \frac{X(t, f, r) - \mu(X)}{\sigma(X) + \varepsilon}, \tag{9}$$

where $\mu(X)$, $\sigma(X)$ and ε are the mean, standard deviation and small positive number, respectively. The small positive number, ε, is stable value added in Eq. 9 to avoid dividing by zero. The mean and standard deviation are calculated as Eq. 10 and Eq. 11:

$$\mu(X) = \frac{1}{TFR}\sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{r=1}^{R} X(t, f, r), \tag{10}$$

$$\sigma(X) = \sqrt{\frac{1}{TFR}\sum_{t=1}^{T}\sum_{f=1}^{F}\sum_{r=1}^{R}(X(t, f, r) - \mu(X))^2}, \tag{11}$$

where T, F, and R denote the number of frames, frequency nodes, and channels, respectively. The standard score layer is utilized in the deep learning models for the monophonic speech separation task in order to accelerate the training process and stabilize the neuron activations. As shown in Eq. 9, the normalized power spectra above the mean value yields positive values, while the spectra below the mean value yields negative values. The covariance shift is handled by this technique during the training phase. The usage of the normalization layer improves the quality of the source separation performance [25], [28], [33].

TABLE 3. List of abbreviations of single-channel speech separation problem.

Abbreviation	Explanation/ definition
GMM	Gaussian mixture model
HMM	Hidden markov model
ICA	Independent component analysis
CASA	Computational auditory scene analysis
NMF	Nonnegative matrix factorization
DNN	Deep neural network
TF	Time-frequency
MSE	Mean square error
DC	Deep clustering
DANet	Deep attractor network
PIT	Permutation invariant training
uPIT	Utterance-level PIT
TasNet	Time-domain audio separation network
DPRNN	Dual-path RNN
DPTNet	Dual-Path Transformer Network
SepFormer	Separation Transformer
SML	Selective Mutual Learning
LSTM	Long-short term memory
BLSTM	Bidirectional long-short term memory
TF	Time-frequency
DFT	Discrete Fourier transform
STFT	Short-time Fourier transform
iSTFT	Inverse short-time Fourier transform
IBM	Ideal binary mask
IRM	Ideal ratio mask
VAD	Voice activity detection
SDR	Source to distortion ratio
SAR	Source to artifacts ratio
SIR	Source to interference ratio
SI-SNRi	Scale-invariant signal-to-noise ratio improvement
SDRi	Source-to-distortion ratio improvement
STOI	Short-time objective intelligibility
PESQ	Perceptual evaluation of speech quality

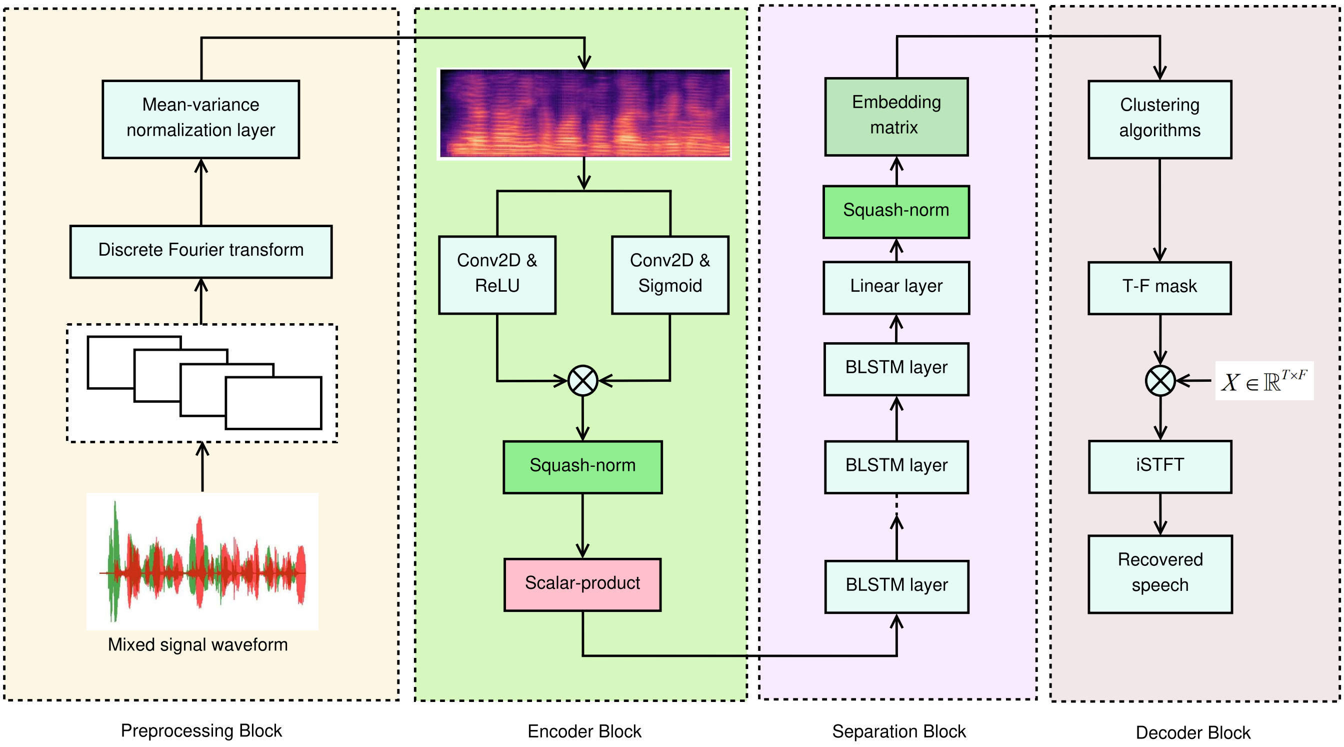


FIGURE 1. The architecture of the proposed ESDC framework uses the node encoder block and the Squash-norm for the input and output embedding vectors. The mixture speech signals are mapped into the high-dimensional embedding vectors. The embedding vectors of TF nodes of the speakers are pulled closer together in the same cluster while being pushed further away in different clusters. The embedding vectors are trained by the model during the training stage, and then they are clustered by the different clustering algorithms to construct the estimated mask during the testing.

B. NODE ENCODER BLOCK

The node encoder block is designed based on the gated convolutional network [51]. The mixed spectral features, $X_i \in \mathbb{R}^{T \times F}$, are fed to two different convolution layers, both with a 5×5 kernel, L feature maps, a stride value of 1, and the same paddings. The sigmoid and ReLU functions are adopted as the non-linearity activation functions for each convolutional stream. These two streams are used to generate the high-dimensional input vectors, $z_i, z_j \in \mathbb{R}^{1 \times L}$, of the input embedding matrix $Z \in \mathbb{R}^{N \times L}$, by element-wise multiplication. Through the convolution layers, the TF nodes are highlighted and mapped onto the input feature vectors. The Squash-norm function is then used to ensure that the small input vectors are reduced to approximately the zero vector, while the large input vectors are scaled below the unit vector.

Instead of using input vectors for the separation block, the input features are replaced by the scalar product, S, of the vectors, S is computed by $z_i z_j = |z_i|\,|z_j|\cos(\widehat{z_i, z_j})$ where $1 \leqslant i, j \leqslant N$, and $N = T \times F$, with the time-frame number T and the frequency-bin number F of the TF nodes. In the view of a vector space, S represents the similarity between the two vectors. If S is positive, the angle between the two vectors is less than ninety degrees. Conversely, when S is negative, the two vectors produce two directions with an angle greater than 90 degrees. On the matter of magnitude, if the absolute value of S is large, it is inferred that the two vectors are also large, or the projection of one vector onto the other is large and vice versa. In other words, the more similar the two input vectors are, the greater their scalar product will be. The high discrimination rate of the feature representation is an important factor for performance improvement in the training stage.

C. SQUASH-NORM OUTPUT VECTORS OF THE SEPARATION BLOCK

Following the main idea of a DC network of the separation block [15], the proposed model trains the scalar product features, and then applies some clustering methods to cluster embedding vectors that generated the estimated TF mask. The scalar product features, S, are used as the input features for the separation block. The scalar product features are trained by the bidirectional long-short term memory (BLSTM) when all of the hidden states of the BLSTM unit are calculated according to the Markov technique. The output products of the BLSTM network are the deep-dimensional vectors, $v_i \in \mathbb{R}^{1 \times D}$, with $1 \leqslant d \leqslant D$, which are mapped from the TF node. These vectors are used to create the deep-dimensional embedding matrix $V = [v_1, v_2, \ldots, v_N]$. The mathematical non-linear function [52] is utilized to normalize the deep-dimensional vector v_i of the matrix $V \in \mathbb{R}^{N \times D}$, and is calculated as Eq 12:

$$u_i = \frac{||v_i||^2}{1 + ||v_i||^2} \frac{v_i}{||v_i||}, \tag{12}$$

where $u_i \in \mathbb{R}^{1\times D}$ denotes the normalized deep-dimensional vector of the normalized deep-dimensional matrix $U \in \mathbb{R}^{N\times D}$. The deep dimension value, D, is the number of elements $v_{i,d} \in \mathbb{R}^{1\times 1}$ in the deep-dimensional vector. The magnitude of the deep-dimensional vector, v_i, is calculated by $||v_i|| = \sqrt{\sum_{d=1}^{D} v_{i,d}^2}$. According to the non-linear function, the magnitude, $||u_i||$, of the normalized deep-dimensional vector u_i is calculated as follows:

$$||u_i|| = \frac{||v_i||^2}{1+||v_i||^2}. \tag{13}$$

From Eq. (13), the magnitude of the normalized deep-dimensional vectors will be in the range from zero to one. The normalized vector u_i becomes close to the zero vector when the vector v_i is a vector with small magnitude. Conversely, this normalized vector becomes close to the unit vector if the vector v_i yields large magnitude. Eq. (12) shows the definition of the Squash-norm that enhances the discriminative learning of the normalized deep-dimensional vectors into disjoint clusters. The labels are represented by the class indicator vectors, $y_i, y_j \in \mathbb{R}^{1\times C}$, of the label ideal ratio mask $Y \in \mathbb{R}^{N\times C}$ for each TF node, where C numbers of single-channel sources in the mixture. The loss function, $L(u)$, is used to optimize the feature representation which means minimum for the difference between the scalar product, $u_iu_j = |u_i|\,|u_j|\,\cos(\widehat{u_i, u_j})$, of the normalized deep-dimensional vector in the normalized matrix $U = [u_1, u_2, \ldots, u_N]$ and the scalar product, $y_iy_j = |y_i|\,|y_j|\,\cos(\widehat{y_i, y_j})$, of the indicator vectors in the label ideal ratio (or label binary ideal) mask $Y = [y_1, y_2, \ldots, y_N]$, and is calculated as Eq. 14:

$$\begin{aligned} L(u) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(u_iu_j - y_iy_j\right)^2 \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\sum_{d=1}^{D} u_{i,d}u_{j,d} - \sum_{c=1}^{C} y_{i,c}y_{j,c}\right)^2. \end{aligned} \tag{14}$$

To reduce the computational consumption, the issue is considered as low-rank for the sparse point Y. Therefore, the loss function is replaced by

$$\begin{aligned} L(u) = &\sum_{d,k=1}^{D}\left(\sum_{i=j=1}^{N} u_{i,d}u_{j,k}\right)^2 + \sum_{c,q=1}^{C}\left(\sum_{i=j=1}^{N} y_{i,c}y_{j,q}\right)^2 \\ &-2\sum_{d=1}^{D}\sum_{c=1}^{C}\left(\sum_{i=j=1}^{N} u_{i,d}y_{j,c}\right)^2, \end{aligned} \tag{15}$$

where $u_{i,d}$, $u_{j,k}$ are the normalized embedding elements and $y_{i,c}$, $y_{j,q}$ are the elements of the indicator vectors. The scalar product of the vectors represents the relationship between the output vectors. In other words, the more similar the pairs of the output vectors are, the greater their scalar products will be. Hence, the embedding vectors in the normalized embedding space are pulled closer together in the same cluster while being pushed further away in different clusters. Both the output dimension mismatch and permutation issues are able to be solved by the loss function and the clustering stage on the normalized vector.

In the case of $\ddot{u}_i$ is unit-norm of v_i, the loss function, $L(\ddot{u})$, is used to optimize the feature representation which means to minimize the difference between the *cosine* of the normalized output vectors and the *cosine* of the indicator vectors, and calculated as follows:

$$L(\ddot{u}) = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\cos(\widehat{\ddot{u}_i, \ddot{u}_j}) - \cos(\widehat{y_i, y_j})\right)^2. \tag{16}$$

The separated method is different from the end-to-end source separation methods. The estimated signal recovered by this method will be described in the next section, decoder block.

D. DECODER BLOCK FOR WAVEFORM RECONSTRUCTION

Visualization of the decoder block used for waveform reconstruction is illustrated in Fig. 2. The separation block produces the normalized embedding vectors during the training stage. The vectors, u_i, of the normalized embedding matrix U are separated into the disjoint clusters when using the different clustering algorithms in the testing stage.

The optimization problem guides a clustering process, e.g., K-means. Given a set of vectors u_i, K-means aims to group the N vectors into C clusters of the vectors. This is achieved by minimizing the cost function:

$$\begin{aligned} &\min \sum_{i=1}^{N}\sum_{c=1}^{C} \tau_{ic}||u_i - m_c||_2^2 \\ &\tau_{ic} \in \{0,1\}\ \forall i, c;\ \sum_{c=0}^{C}\tau_{ic} = 1 \end{aligned} \tag{17}$$

where $m_c \in \mathbb{R}^{1\times D}$ is the centroid of the c^{th} cluster. τ_{ic} is element of a label vector τ_i and denotes the c^{th} cluster's membership assignment.

The estimated TF mask, $\widehat{M}_c(t,f) \in \mathbb{R}$, of each speech source recreates the estimated speech sources in the mixed signal as previously shown in Fig. 1, and is written as Eq. 18:

$$\widehat{M}_c(t,f) = ClusteringAlgorithm\left(\{u_i\}_{i=1}^{N}, C\right). \tag{18}$$

The estimated signal, $\widehat{S}_c(t,f) \in \mathbb{C}$, of each source signal is calculated by $\widehat{S}_c(t,f) = \widehat{M}_c(t,f) \odot X(t,f)$. The estimated speech signal $\widehat{S}_c(t,f)$ in the TF domain is utilized to rehabilitate each estimated speech signal $\widehat{s}_c(n) \in \mathbb{R}$ in the time domain by iSTFT.

VI. EXPERIMENTS

A. EXPERIMENTAL SETUP

1) FEATURES

All experiments were performed on speech mixtures generated from the TSP speech corpus [53] and the TIMIT speech

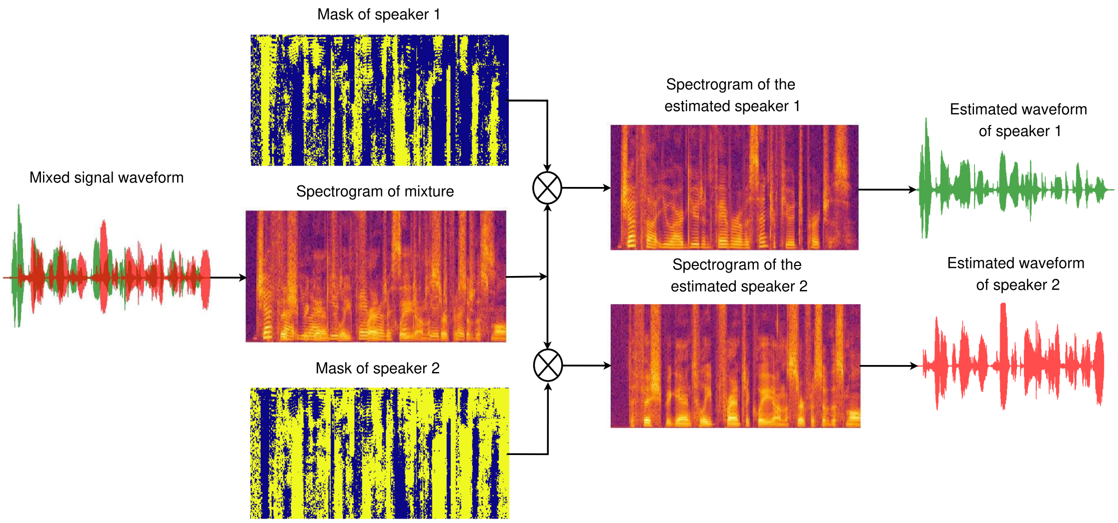

Mask of speaker 1
Mixed signal waveform
Spectrogram of mixture
Mask of speaker 2
Spectrogram of the estimated speaker 1
Spectrogram of the estimated speaker 2
Estimated waveform of speaker 1
Estimated waveform of speaker 2

FIGURE 2. Visualization of the decoder block describes the estimation masks. These estimation masks perform element-wise multiplication with the mixed signal to recreate the spectrograms of the first and the second speakers in the TF domain. Then, these spectrograms are transformed into time-domain waveforms by iSTFT.

corpus [54] for acoustic-phonetic data. The TSP dataset contains the recordings of 24 speakers with 1444 utterances with an average length of 2.372 s, and almost half male and half female. The TIMIT dataset comprises of broadband recordings of 630 speakers in eight major dialects in American English, each with 16kHz speech waveform files and ten phonetically rich sentences per speaker. In the case of different genders, we select four speakers FA, FB, MC, and MD from the TSP speech corpus for experiments, mixing together 60 voices of the speaker. The data is divided into groups 80% for training, 10% for development, and 10% for evaluation. The utterance datasets are constructed from the TIMIT corpus. The training and testing directories are divided into the original TIMIT corpus with the training, development, and evaluation datasets. Both TSP and TIMIT datasets are mixed between -5 and 5 dB signal-to-noise ratio. In the case of the same gender, we only select the same male or female speech from the original TIMIT corpus in the training and testing directories. The length of the input mixed voice is the shortest speech in the component sources. The log-scaled magnitude spectrograms of the STFT using input features are down sampled from 16 kHz to 8kHz to reduce the computational consumption. The window length is 32 ms using a Hann window with a hop length of 8 ms. These input features have a 256-point DFT signal. A voice activity detection (VAD) threshold is applied to each spectral frame, removing inactive TF bins during the computation of the objective function. Only TF bins with magnitude greater than VAD were used during training (VAD threshold was set to 40 dB). VAD ensures that the model does not assign vectors to inactive TF bins. In this way, the computational cost is significantly reduced.

2) OBJECTIVE EVALUATION METRICS

For the objective measures of performance, we use several criteria, including scale-invariant signal-to-noise ratio improvement (SI-SNRi), source-to-distortion ratio improvement (SDRi), short-time objective intelligibility (STOI) [55], and perceptual evaluation of speech quality (PESQ) [56]. In addition, source to distortion ratio (SDR), source to artifacts ratio (SAR), and source to interference ratio (SIR) in the BSS-EVAL toolbox [57] are used for comparison with the other methods. Higher index values represent better separation quality.

TABLE 4. The values of feature and parameters in single-channel speech separation.

Parameter	Value
Window length	32 ms
Hop length	8 ms
Window	Hann
Sampling frequency	8 kHz
DFT size	256
Numbers of single-channel sources	2, 3
Embedding dimension value	40
Feature map	16
Number of BLSTM units	600
Number of BLSTM layers	4
Batch size	32
Number of epochs for training	120

STOI represents a quantified index of speech intelligibility, ranging from 0 to 1. It revealed the correlation of voice intelligibility in hearing tests.

PESQ is a quantitative estimate of the source separation in the ranges in $[-0.5, 4.5]$, and uses cognitive modeling to measure interference between the pure voice and the estimated voice. PESQ is calculated as follows:

$$PESQ = 4.5 - 0.1d_{SYM} - 0.0309d_{ASYM}, \tag{19}$$

where d_{SYM} and d_{ASYM} denote symmetric disturbance and asymmetric disturbance, respectively.

The SIR demonstrates the ability to reduce interference. The SIR is determined only on the basis of interference from other signals, with noise being ignored. The SAR analyses variations in the signal, such as brief audio spikes lasting a few milliseconds or less. The SDR evaluates the cumulative distortion of these various effects compared to the original signal, and is therefore frequently used as a general measurement of how effectively the separation work. SDR, SIR, and SAR in the BSS-EVAL toolbox are respectively defined as:

$$\mathrm{SDR}(s, \hat{s}) = 10\log_{10}\frac{||s_{\mathrm{target}}||^2}{||e_{\mathrm{interf}} + e_{\mathrm{noise}} + e_{\mathrm{artif}}||^2}, \quad (20)$$

$$\mathrm{SIR}(s, \hat{s}) = 10\log_{10}\frac{||s_{\mathrm{target}}||^2}{||e_{\mathrm{interf}}||^2}, \quad (21)$$

$$\mathrm{SAR}(s, \hat{s}) = 10\log_{10}\frac{||s_{\mathrm{target}} + e_{\mathrm{interf}} + e_{\mathrm{noise}}||^2}{||e_{\mathrm{artif}}||^2}, \quad (22)$$

where e_{interf}, e_{noise}, and e_{artif} are the interferences, noise and artifacts error terms, respectively. s, $\hat{s}$ and x are a reference clean source, an estimated source, and a mixture in the time domain, respectively. As can be seen, the evaluation metrics are very powerful in analyzing the results of an algorithm. However, the noise signals are not always known so we don't compute e_{noise} and e_{artif}. This is SI-SNR used as SDR. SI-SNR is utilized as an objective measure and is defined as follows:

$$s_{\mathrm{target}} = \frac{\langle \hat{s}, s\rangle s}{||s||^2}, \quad (23)$$

$$e_{\mathrm{noise}} = \hat{s} - s_{\mathrm{target}}, \quad (24)$$

$$\mathrm{SI\text{-}SNR}(s, \hat{s}) = 10\log_{10}\frac{||s_{\mathrm{target}}||^2}{||e_{\mathrm{noise}}||^2}. \quad (25)$$

The SI-SNRi and SDRi metrics used in [17], [22], [24], and [28], computed as follows:

$$\mathrm{SI\text{-}SNRi}(s, \hat{s}, x) = \mathrm{SI\text{-}SNR}(s, \hat{s}) - \mathrm{SI\text{-}SNR}(s, x) \quad (26)$$

$$\mathrm{SDRi}(s, \hat{s}, x) = \mathrm{SDR}(s, \hat{s}) - \mathrm{SDR}(s, x) \quad (27)$$

3) SYSTEM ARCHITECTURE AND REGULARIZATION

The training network in the ESDC model is built by Tensorflow. Both the feature extraction and signal reconstruction have used the Librosa library and Signal library in Scipy. Referencing to [15]–[18], we adopt various hyper-parameters to adjust the number of units and layers for robustness. In Fig. 1, the ESDC model is constructed by four BLSTM layers with 600 hidden nodes in each layer, the number of clusters is $C = 2$ and 3, the feature map is $L = 16$, the dimension is $D = 40$ with a sigmoid activation function before the output layer. Different clustering algorithms are used to construct the estimated mask. We choose the Adam algorithm [58] as the optimizer with settings of $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 1e-8$, the initial learning rate $\eta = 0.0001$ and the batch size of 32. The alternative learning rate iteratively halves the previous learning rate while the objective function is not reduced by more than 4 epochs on the validation set. The dropout technique is used to prevent over-fitting, the input dropout is 50% and the output dropout is 20%. The training stage using the early stopping technique automatically stops and saves the alternative weights when the objective function does not decrease on the development set for more than 5 epochs. The number of epochs for training is 120.

B. COMPARISONS AND EXPERIMENTAL RESULTS

In the experiments, the proposed models are evaluated on the monaural source separation tasks with two cases, different genders on the TIMIT and TSP datasets and the same gender on the TIMIT dataset. We evaluate the obtained results of the speech separation task by testing each system block separately and divide the tests into five parts. In the first part, we evaluate the separation results in comparison with other methods on the TIMIT dataset. In the second part, we compare the proposed models to the experimental results. In the third part, the clustering algorithms are studied through the proposed ESDC model. In the fourth part, we investigate the types of TF masks to construct the indicator vectors and the output activation functions. In the last part, we evaluate the similarity matrix in the proposed cases and the spectrograms of the best model. In this work, we evaluate the speech separation performances when training many different models on the TIMIT and TSP datasets. We use the DC model which consists of four BLSTM layers with 600 hidden units for each layer and a fully connected layer. The input features, $X_i \in \mathbb{R}^{T\times F}$, are used in the DC model, and the output embedding vectors are created by a fully connected layer. The output vectors are normalized by the unit-norm. In the Squash-norm DC model (SDC), we use the Squash-norm instead of the unit-norm. The Squash-norm is used to normalize the output embedding vectors, and the SDC model directly trains the normalized spectra $X_i \in \mathbb{R}^{T\times F}$. In the ESDC model, we combine the node encoder and the SDC model. Two different convolutional layers, the sigmoid and ReLU functions are utilized to construct the input vectors. These input vectors are normalized by the Squash norm. Then a scalar product, $z_i z_j = |z_i|\,|z_j|\cos(\widehat{z_i, z_j})$, of the vectors is used to replace the input features for SDC.

1) OBJECTIVE EVALUATION

In Table 5, we compare the separation performances obtained by the proposed ESDC model and the other models on SDR, SAR, and SIR metrics in the different gender speech case, including CNN [59], DF-DNN [60], RNN [61], DRNNs [62], and GDC [63]. In this case, the ESDC model in Table 5 achieves an improved performance compared to the other methods. As shown in the table, the proposed ESDC model increase 3.75 – 6.46 dB SDR, 8.91 – 11.17 dB SIR, and 3.64 — 6.3 dB SAR compare with the others.

We utilize DC as the framework baseline to evaluate the proposed models. To further analyze the performance of the proposed model, we conduct the experiments on the separation performances in several aspects. In these experiments,

TABLE 5. The values in dB of the proposed ESDC model using the indicator vectors from *IRM* and K-means algorithm for the estimated mask compare to the values of the others on the TIMIT dataset.

Model	SDR dB	SIR dB	SAR dB
CNN [59]	6.60	9.50	10.20
DF-DNN [60]	6.39	9.72	9.89
RNN [61]	7.41	11.76	7.54
DRNNs [62]	7.74	-	-
GDC [63]	9.1	-	-
ESDC	**12.85**	**20.67**	**13.84**

the proposed models present two cases including SDC and ESDC. Table 6 presents the experiment results of DC, SDC, and ESDC on the TIMIT and TSP datasets. It can be observed that the separation results on the TSP dataset are always lower than the TIMIT dataset on all three trained models. This demonstrates that deep learning models give better results when the training data is larger. In Table 6, the quality values of SDR, SDRi, SI-SNRi, STOI, and PESQ of SDC are respectively 1.09 dB, 1.09 dB, 1.11 dB, 2%, and 0.21 higher respectively, compared to DC on the TIMIT dataset. These separation performances are similar on the TSP dataset where the evaluation metrics increase by 1.78 dB SDR, 1.91 dB SDRi, 2.03 dB SI-SNRi, 5% STOI, and 0.35 PESQ gain, respectively. ESDC provides the best separation performance, where SDR, SDRi, SI-SNRi, STOI, and PESQ metrics are respectively 1.27 dB, 1.28 dB, 1.3 dB, 3%, and 0.29 higher than DC, and respectively 0.18 dB, 0.19 dB, 0.19 dB, 1%, and 0.08 higher than SDC on the TIMIT dataset. These values of ESDC on the TSP dataset increase respectively 2.09 dB SDR, 2.21 dB SDRi, 2.44 dB SI-SDRi, 6% STOI, and 0.39 PESQ compare to DC and respectively 0.31 dB SDR, 0.3 dB SDRi, 0.41 dB SI-SDRi, 1% STOI, and 0.04 PESQ higher than SDC. However, the usage of the node decoder of the ESDC model results in a significant increase in the number of the parameters to 114,893,992, while the number of parameters of DC and SDC is smaller and equal. It can be seen that SDC achieves better separation performance than DC because the Squash-norm shrinks or expands the output vectors of the separation block. On the other hand, The performance of the ESDC model increases slightly compared to SDC when we add the node encoder for ESDC.

In Table 7, we compare the proposed ESDC method with DC and SDC on the TIMIT dataset with the case of the same gender. The proposed ESDC approach has the best performance for the same gender pairs. The proposed ESDC approach outperforms DC and, produces slightly better separation performances than SDC. However, the separation performances of the same gender in Table 7 are much lower than the different gender in Table 6. The experimental results in Tables 6 and 7 are consistent with the previous observations in [64] and [65] where the same gender speech separation from the mixed signal is always a difficult issue due to the similar frequency range from the same gender.

The proposed models consist of two phases. In the first phase, we utilize the scalar product of the indicator vectors to estimate the scalar product of the high-dimensional output vectors through supervised learning. In the second phase, we study these output vectors through unsupervised learning using clustering algorithms. According to Table 8, we investigate the separation performances of the different clustering algorithms on the proposed ESDC model on the TSP and TIMIT datasets. Gaussian Mixture Models (GMM), Bayesian Gaussian Mixture (BGM), Mini batch K-means (MBK-means), K-means, K-means++, and Spherical K-means (SK-means) algorithms are selected as the clustering algorithms for TF mask estimation. In the experiments, both GMM and BGM have poor separation performances on embedding space while rest of clustering algorithms have achieved similar performances. The K-means algorithm is less stable for the initialization of the centroids and also less efficient due to the presence of hyperpolynomials in the input. The K-means++ algorithm [66] ensures a better initialization of the centroids, while the MBK-means in [67] is the fastest among the other algorithms. The SK-means clustering method in [68] normalizes the data by using cosine similarity for cluster assignment. At the end of each maximization step, the estimated cluster centroids are mapped onto the unit sphere. The SK-means clustering is superior to K-means on the directional data, both in performance and speed. The comparisons show that the SK-means clustering algorithm outperforms other clustering algorithms.

In Table 9, we perform an ablation experiment on the effect of the Squash-norm on the input vector. From the experiments, we see that the separation performance is slightly enhanced. This proves that Squash-norm shrinks or expands the input vectors, which leads to obtain the discriminative scalar product. The high distinction ratio of the scalar product feature is an important factor for performance improvement.

As shown in Table 6, an important issue is that, the more discriminative the output embedding vectors, the more accurate the estimated mask. This significantly improves the separation performance (DC vs SDC). For input vectors, the more similar the input embedding vectors are, the greater their scalar product is. However, this did not significantly improve the separation performance in Table 9 (Squash-norm vs Non-Squash-norm for the input embedding vectors). Furthermore, we investigate different clustering techniques for output embedding vectors in Table 8. However, the clustering algorithms have little impact on performance. The separation performance mainly depends on the training of the output embedding vectors in Table 6. This proves that the performance mainly depends on discriminative learning of the embedding vectors.

In the Table 10, the backbone network is changed from BLSTM to LSTM. In this case, ESDC is a causal speech separation system suitable for real-time speech separation applications. The same parameters of the BLSTM network (non-causal speech separation system) are used for the LSTM network in this ablation study. We can observe that the

TABLE 6. Separation performance of the proposed models with the indicator vectors from IRM and K-means algorithm on the TSP and TIMIT datasets for the different gender.

Dataset	Model	Parameter	SDR (dB)	SDRi (dB)	SI-SNRi (dB)	STOI	PESQ
TSP	DC	35,635,560	8.10	7.41	6.73	0.82	1.64
	SDC	35,635,560	9.88	9.32	8.76	0.87	1.99
	ESDC	114,893,992	10.19	9.62	9.17	0.88	2.03
TIMIT	DC	35,635,560	11.58	11.15	10.73	0.88	2.15
	SDC	35,635,560	12.67	12.24	11.84	0.90	2.36
	ESDC	**114,893,992**	**12.85**	**12.43**	**12.03**	**0.91**	**2.44**

TABLE 7. Separation performance of the proposed models with the indicator vectors from IRM and K-means on the TIMIT datasets for the same gender.

Gender	Model	SDR (dB)	SDRi (dB)	SI-SNRi (dB)	STOI	PESQ
Male-Male	DC	7.11	6.67	6.62	0.78	1.54
	SDC	8.08	7.63	7.15	0.79	1.65
	ESDC	8.25	7.66	7.24	0.80	1.67
Female-Female	DC	6.79	6.41	5.68	0.75	1.62
	SDC	8.94	8.58	7.96	0.80	1.78
	ESDC	9.50	9.15	8.54	0.82	1.81

TABLE 8. The separation performance of the different clustering algorithms for the output embedding vectors of the proposed ESDC model with the indicator vectors from IRM on the TSP and TIMIT dataset.

Model	Dataset	Clustering algorithms	SDR (dB)	SDRi (dB)	SI-SNRi (dB)	STOI	PESQ
ESDC	TSP	GMM	5.98	5.42	4.51	0.75	1.60
		BGM	6.03	5.46	4.55	0.76	1.61
		MBK-means	10.17	9.61	9.17	0.88	2.03
		K-means	10.19	9.62	9.17	0.88	2.03
		K-means ++	10.19	9.63	9.18	0.88	2.03
		SK-means	10.25	9.71	9.27	0.89	2.03
ESDC	TIMIT	GMM	3.02	2.60	1.48	0.69	1.58
		BGM	3.27	2.85	1.74	0.70	1.59
		MBK-means	12.85	12.42	12.03	0.90	2.44
		K-means	12.85	12.43	12.03	0.91	2.44
		K-means ++	12.87	12.44	12.04	0.91	2.46
		SK-means	**12.91**	**12.48**	**12.08**	**0.91**	**2.46**

separation performance decreases by 1.05 dB SDR, 1.13 dB SIR, 0.93 dB SAR, 0.98 dB SDRi, and 0.93 dB SI-SNRi, respectively while replacing BLSTM to LSTM. Despite the reduced separation performance, this causal system is suitable a variety of applications on real-time devices.

To further investigate the proposed model, we observed the effect of the indicator vectors constructed from IBM, IRM, and PSM in comparison between groups including the first, second grouped columns, the third, fourth grouped columns, and the fifth, sixth grouped columns. The effect of the output activation functions is examined in the first, third, and fifth grouped columns compared to the second, fourth, and sixth grouped columns in Fig. 3. In this work, we use the scalar product of the indicator vectors as the training target. The IBM, IRM, and PSM are used to construct the indicator vectors. In this case, we find that the scalar product objective of the PSM-based indicator vector achieves the highest separation performance, while the scalar product objective of the IRM-based indicator vector achieves higher performance than the IBM-based indicator vector. The scalar product objective based on the IBM-based indicator vector is more sensitive to estimation error than the IRM-based indicator vector. The scalar product objective of PSM-based indicator vectors adds phase information, which achieves higher performance than IBM-based and IRM-based indicator vectors. The performance difference using scalar product targets based on IBM-based, IRM-based, and PSM-based metrics vectors is negligible. In our method, the effect of the

TABLE 9. The values in dB of the proposed ESDC model with non-Squash encoder and Squash encoder using the K-means algorithm on the TIMIT dataset.

Encoder	SDR	SIR	SAR	SDRi	SI-SNRi
Non-Squash	12.74	19.72	13.68	12.31	12.03
Squash	12.85	20.67	13.84	12.43	12.03

TABLE 10. The values in dB of the proposed ESDC model with LSTM and BLSTM using the K-means algorithm on the TIMIT dataset.

Model	Backbone	SDR	SIR	SAR	SDRi	SI-SNRi
ESDC	LSTM	11.80	19.54	12.91	11.36	11.10
ESDC	BLSTM	12.85	20.67	13.84	12.43	12.03

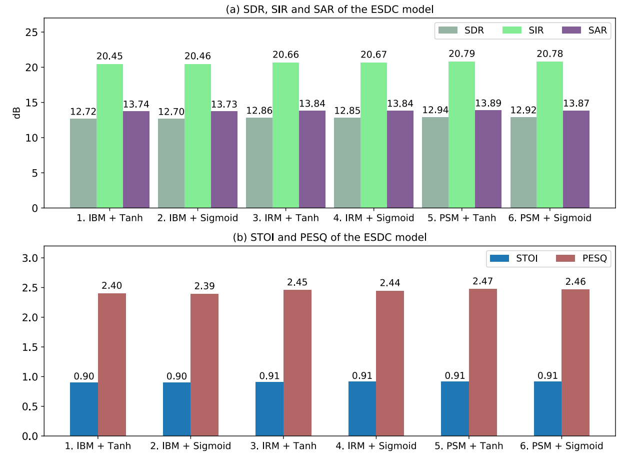


FIGURE 3. The results of speech separation performance of the ESDC model with K-means on the TIMIT dataset using different activation functions and the indicator vectors from the different reference masks.

activation function is negligible, as shown in Fig. 3. The differences between the six cases are insignificant because our method relies heavily on the training phase of the embedding vector and the clustering algorithm that estimates the mask.

The full neural network-based methods achieve impressive performance in scenarios with a fixed number of sources. Each source belongs to a discriminative signal class, such as speech and music. However, the number of classes and sources is arbitrary, making fully neural network-based approaches unsuitable. In contrast, a deep clustering strategy comparable to deep learning network-based spectral clustering assigns an embedding vector to each TF bin of the spectral map. Then, the vectors are clustered by the clustering algorithm to construct the estimated mask. In Table. 11, we train the ESDC model with two-speaker mixtures and three-speaker mixtures. The two-speaker mixed model performed well in tests with two and three speakers. In particular, the three-speaker model achieves impressive performance in both cases. This proves that the ESDC model with deep clustering achieves good results in both with fixed and arbitrary source cases.

TABLE 11. SDR (dB) of the ESDC model in the case of training with different numbers of speakers on the TIMIT dataset.

Model	Training dataset	Test dataset	
		Two speakers	Three speakers
ESDC	Two speakers	12.85	4.36
	Three speakers	10.21	8.71

2) SPECTROGRAM ANALYSIS AND SIMILAR MATRIX ANALYSIS

Observing the spectrograms in Fig. 4, and 5, the proposed ESDC model performs source separation of gender-specific speech on the TSP and TIMIT datasets. Compared with the original spectrogram, the recovered spectrogram has a high similarity. Whereas in Fig. 6, and 7, we perform the same gender speech separation. Some spectral regions confused and overlapped between the two speeches. We know that speech separation of the same gender is more difficult than the case of different genders. Since the pitch and vocal tract of the voices are in the same range, separation from mixed speech of same-sex speakers can be extremely challenging.

In this experiment, we construct the similarity matrices in Fig. 8(a), (b), and (c). These similarity matrices are the scalar product of output vectors of DC, SDC, and ESDC in the training stage, respectively. Each element of the similarity matrix is the scalar product of the two output vectors. The similarity matrix reflects the relationship between TF nodes and densely connected information. The blue areas have higher similarity than the white areas. A segment of speech data is randomly chosen for observation. In this case, we use the frame in the female speech {`./TSP/FA_TEST/FA05_09.wav`} and the frame in the male speech {`./TSP/MB_TEST/MB12_10.wav`} in the TSP dataset. Observed on the color area of each similarity matrix, the blue areas are small and the contrast of the elements on the similarity matrix in Fig. 8(a) is low. Therefore, it is difficult to determine the relationships of the elements in the matrix. In Fig. 8(b) and (c), we use the Squash-norm which shrinks the small vectors and dilates the large ones. It increases the contrast of the scalar product of the vectors. Therefore, the blue areas are enlarged, similar to the high contrast of elements on the matrix. In the experiments, when the training model is combined with the node decoder and the Squash-norm function, the contrast of the similarity matrix is most evident in Fig. 8(c).

The similarity on the diagonal of the matrix is the highest in Fig. 8(a) with a value of 1 but it is only the highest value of Fig. 8(b) and (c). To explain this, the values on the diagonal are the scalar product of the same vector. The unit normalization is used in Fig. 8(a) so that the values on the diagonal are always 1, while these values in Fig. 8(b) and (c) are normalized by the Squash-norm function. In other words, the TF nodes of a speech frame in the blue areas of the similarity matrix may belong to the same speech source for

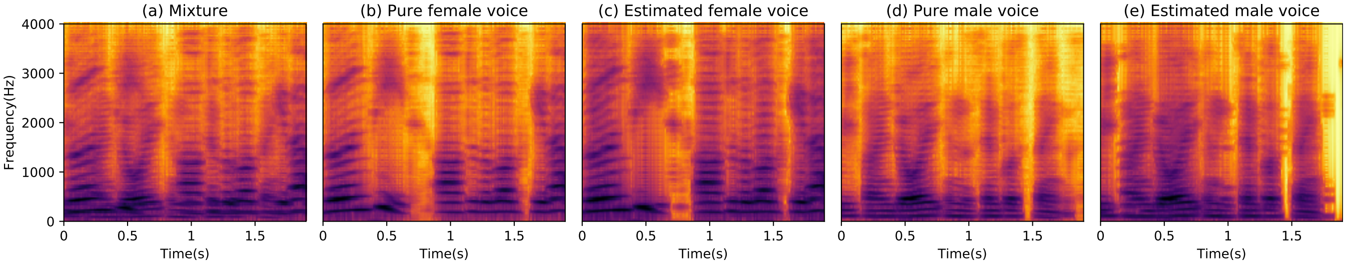


FIGURE 4. A voice separation example of the ESDC model on the TSP dataset. Fig. 4(a) is the mixed spectrum of the pure-female voice and the pure-male voice. Fig. 4(b) and (c) are the spectra of the pure-female voice and the estimated-female voice, respectively. Fig. 4(d) and (e) are the spectra of the pure-male voice and the estimated-male voice, respectively.

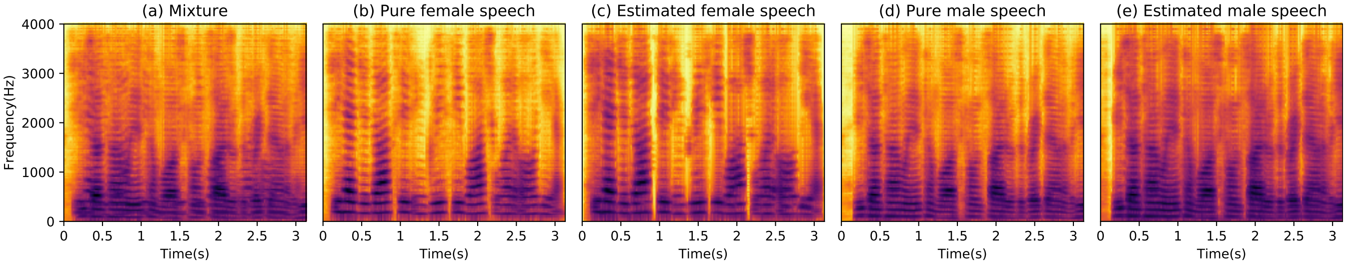


FIGURE 5. The spectrograms of the different-gender speech separation example of the ESDC model on the TIMIT dataset. Fig. 5(a) shows the mixed spectrum of the pure-female speech and the pure-male speech. Fig. 5(b) and (c) are the spectra of the pure-female speech and the estimated-female speech, respectively. Fig. 5(d) and (e) are the spectra of the pure-male speech and the estimated-male speech, respectively.

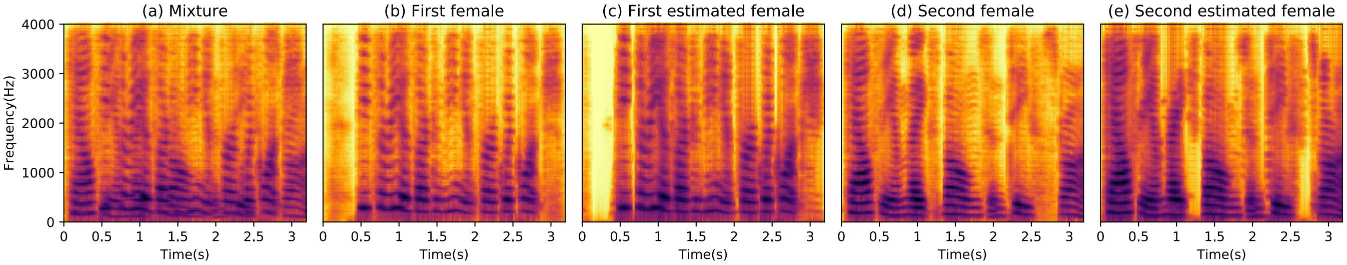


FIGURE 6. Illustration of spectrograms for separating the first female and second female utterances from the mixed utterance of the ESDC model on the TIMIT dataset. Fig. 6(a) is the mixture of the pure speech of the first female and the pure speech of the second female. Fig. 6(b) and (c) show the spectra of the pure speech of the first female and the estimated speech of the first female, respectively. Fig. 6(d) and (e) show the spectra of the pure speech of the second female and the estimated speech of the second female, respectively.

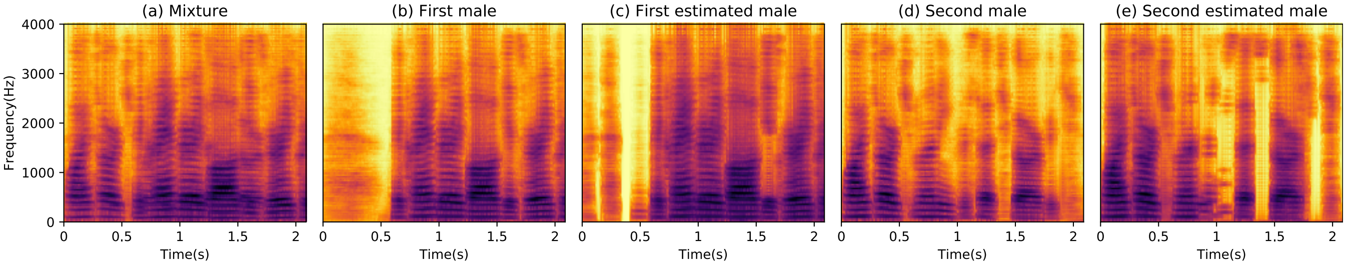


FIGURE 7. Illustration of spectrograms for separating the first male and second male utterances from the mixed utterance of the ESDC model on the TIMIT dataset. Fig. 7(a) is the mixture of the pure speech of the first male and the pure speech of the second male. Fig. 7(b) and (c) are the spectra of the pure speech of the first male and the estimated speech of the first male, respectively. Fig. 7(d) and (e) are the spectra of the pure speech of the second male and the estimated speech of the second male, respectively.

the high similarity while the TF nodes of a speech frame in the white areas of the similarity matrix may belong to different speech sources with low similarity. Therefore, we can further improve the separating performance of the embedding vector of matrix based on the Squash-norm constructed by the speech similarity matrix, and obtain embedding with structural features that are beneficial for estimating TF masks.

3) SUBJECTIVE EVALUATION

The subjective evaluation is done through a series of blind AB listening tests that determined the method

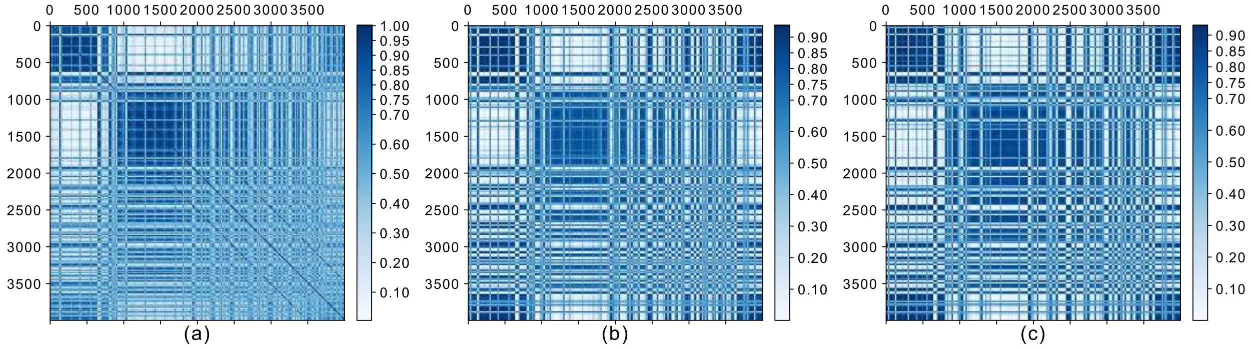


FIGURE 8. Visualization of similarity matrix weights, Fig. 8(a), (b), and (c) are matrix weights of DC, SDC, and ESDC models, respectively. The similarity matrix weights are the output product of the training models. The weights of the matrix show high similarity or low similarity of the output embedding vector.

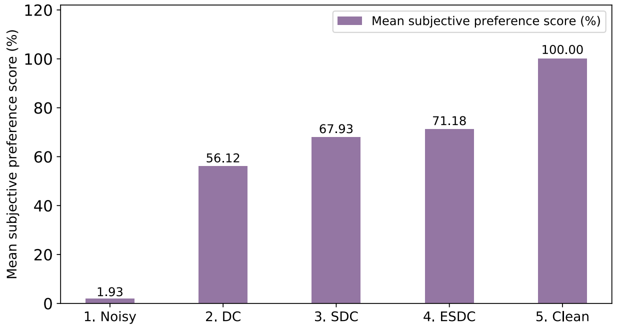


FIGURE 9. The mean subjective preference score (%) comparison of the ESDC, SDC, and DC for the two recordings sp05 and sp07 with 5 dB SNR in the Noizeus corpus.

preference [69], [70]. We performed to create stimuli of sp10 and sp27 from the NOIZEUS corpus at 5 dB SNR. Utterance sp10 and sp27 are male and female speakers, respectively. This experiment was helped by twenty-five listeners. The actual test comprised of thirty stimulus pairs shuffled at a comfortable volume through closed circumaural. For a subjective preference for each stimulus pair, listeners were presented with three choices. The first and second options indicated a preference for the associated stimuli, while the third option indicated a similar liking for both stimuli. Pairwise scoring was used, with the preferred approach receiving a score of 100% and the other receiving a score of 0%. Each method received a 50 % score for a similar preference response.

In addition to other objective evaluations, we also conduct subjective evaluations of the proposed ESDC model. In Fig. 9, the clean speech, ESDC, SDC, DC, and noisy achieve 100%, 71.18%, 67.93%, 56.12%, and 1.93% of the mean subjective preference score, respectively. The subjective evaluations showed that clean speech is always the most preferred, while noisy is the least preferred. The subjective evaluation has shown that our proposed ESDC method is more popular with listeners than other methods. Among the remaining methods, the listeners preferred the SDC method over the DC method. On the AB blind listening tests, we find that the proposed ESDC method achieves the best subjective evaluation among all the tested methods.

VII. CONCLUSION

In this paper, we propose an ESDC framework leveraging Squash-norm and the DC node encoder. The node encoder highlights and enhances scalar product features, while Squash-norm improves the discrimination between high-dimensional vectors in the embedding matrix. When combining node encoder, DC, and Squash-norm, we explore various deep learning architectures, including DC, SDC, and ESDC, to solve the monophonic speech separation problem in the TF domain. Node encoder and Squash-norm significantly improve the performance of the proposed ESDC model. Overall, the proposed ESDC model achieves 1.27 – 2.09 dB SDR, 1.28 – 2.21 dB SDRi, and 1.3 – 2.44 dB SI-SNRi in TSP and TIMIT dataset gains compared to the DC baselines in different gender situations separation performance. Compared with the DC baseline on the TIMIT dataset, our proposed ESDC achieves 1.14 – 2.71 dB SDR, 0.99 – 2.74 dB SDRi, and 0.62 – 2.86 dB SI-SNRi gains in the same-sex case of separation tasks. In all cases, the proposed ESDC model consistently maintains STOI and PESQ higher than the DC baseline. Furthermore, the ESDC method also outperforms the other methods in Table 5. In future research, we will design models with diffusion maps in graph representation learning and graph neural networks to accomplish the task of speech separation, which is expected to well distinguish elements in similarity matrices.

REFERENCES

[1] S. Arberet, A. Ozerov, R. Gribonval, and F. Bimbot, "Blind spectral-GMM estimation for underdetermined instantaneous audio source separation," in *Proc. Int. Conf. Independ. Compon. Anal. Signal Separat.* Berlin, Germany: Springer, 2009, pp. 751–758.

[2] T. Virtanen, "Speech recognition using factorial hidden Markov models for separation in the feature space," in *Proc. Interspeech*, Sep. 2006, pp. 1–4.

[3] S. Choi, A. Cichocki, H. M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Inf. Process. Lett. Rev.*, vol. 6, no. 1, pp. 1–57, 2005.

[4] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, pp. 2067–2079, 2010.

[5] K. Hu and D. Wang, "An unsupervised approach to cochannel speech separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 1, pp. 122–131, Jan. 2013.

[6] F. R. Bach and M. I. Jordan, "Spectral clustering for speech separation," in *Automatic Speech and Speaker Recognition: Large Margin and Kernel Methods*, vol. 369. Hoboken, NJ, USA: Wiley, 2009, pp. 221–253.

[7] G. Hu and D. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 556–562.

[9] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1066–1074, Mar. 2007.

[10] D.-Q. Vu, T.-T.-T. Phung, and J.-C. Wang, "A novel self-knowledge distillation approach with Siamese representation learning for action recognition," in *Proc. Int. Conf. Vis. Commun. Image Process. (VCIP)*, Dec. 2021, pp. 1–5.

[11] D.-Q. Vu, N. Le, and J.-C. Wang, "Teaching yourself: A self-knowledge distillation approach to action recognition," *IEEE Access*, vol. 9, pp. 105711–105723, 2021.

[12] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[13] S. Parveen and P. Green, "Speech enhancement with missing data techniques using recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, p. 733.

[14] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *Proc. IEEE Int. Workshop Electron., Control, Meas., Signals Their Appl. Mechatronics (ECMSM)*, May 2017, pp. 1–5.

[15] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 31–35.

[16] Z.-Q. Wang, J. L. Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 686–690.

[17] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, Sep. 2016, pp. 545–549.

[18] Z.-Q. Wang, J. L. Roux, D. Wang, and J. Hershey, "End-to-end speech separation with unfolded iterative phase reconstruction," in *Proc. Interspeech*, Sep. 2018, pp. 2708–2712.

[19] Y. Jin, C. Tang, Q. Liu, and Y. Wang, "Multi-head self-attention-based deep clustering for single-channel speech separation," *IEEE Access*, vol. 8, pp. 100013–100021, 2020.

[20] L. Li and H. Kameoka, "Deep clustering with gated convolutional networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 16–20.

[21] H. M. Tan and J.-C. Wang, "Single channel speech separation using enhanced learning on embedding features," in *Proc. IEEE 10th Global Conf. Consum. Electron. (GCCE)*, Oct. 2021, pp. 430–431.

[22] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 246–250.

[23] D. Yu, M. Kolbaek, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 241–245.

[24] M. Kolbaek, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[25] Y. Liu and D. Wang, "Causal deep CASA for monaural talker-independent speaker separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2109–2118, 2020.

[26] Y. Zhao, D. Wang, B. Xu, and T. Zhang, "Monaural speech dereverberation using temporal convolutional networks with self attention," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1598–1607, 2020.

[27] E. W. Healy, H. Taherian, E. M. Johnson, and D. Wang, "A causal and talker-independent speaker separation/dereverberation deep learning algorithm: Cost associated with conversion to real-time capable operation," *J. Acoust. Soc. Amer.*, vol. 150, no. 5, pp. 3976–3986, Nov. 2021.

[28] Y. Luo and N. Mesgarani, "TaSNet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 696–700.

[29] N. Zeghidour and D. Grangier, "Wavesplit: End-to-End speech separation by speaker clustering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 2840–2849, 2021.

[30] Z. Shi, H. Lin, L. Liu, R. Liu, and J. Han, "Furcanet: An end-to-end deep gated convolutional, long short-term memory, deep neural networks for single channel speech separation," in *Proc. Int. Conf. Multimedia Modeling*, 2020, pp. 653–665.

[31] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: Multi-scale neural network for end-to-end audio source separation," in *Proc. 19th Int. Soc. Music Inf. Retr. Conf., (ISMIR)*, 2018, pp. 334–340.

[32] Z. Shi, R. Liu, and J. Han, "LaFurca: Iterative context-aware end-to-end monaural speech separation based on dual-path deep parallel inter-intra bi-LSTM with attention," 2020, *arXiv:2001.08998*.

[33] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 46–50.

[34] E. Nachmani, Y. Adi, and L. Wolf, "Voice separation with an unknown number of multiple speakers," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7164–7175.

[35] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Proc. Interspeech*, Oct. 2020, pp. 1–5.

[36] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 21–25.

[37] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 5759–5763.

[38] H. M. Tan, D.-Q. Vu, C.-T. Lee, Y.-H. Li, and J.-C. Wang, "Selective mutual learning: An efficient approach for single channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 3678–3682.

[39] Y. Luo, C. Han, and N. Mesgarani, "Ultra-lightweight speech separation via group communication," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 16–20.

[40] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 61–65.

[41] S. Roweis, "One microphone source separation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 13, 2000, pp. 763–769.

[42] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[43] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, p. 817.

[44] A. M. Reddy and B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in *Proc. Interspeech*, Oct. 2004, pp. 1–4.

[45] A. N. Deoras and A. Hasegawa-Johnson, "A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2004, p. 861.

[46] M. N. Schmidt and R. K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Proc. Interspeech*, Sep. 2006, pp. 1–4.

[47] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 3734–3738.

[48] E. M. Grais, G. Roma, A. J. Simpson, and M. D. Plumbley, "Single-channel audio source separation using deep neural network ensembles," in *Audio Engineering Society Convention*. New York, NY, USA: Audio Engineering Society, 2016.

[49] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[50] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7092–7096.

[51] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 933–941.

[52] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.

[53] P. Kabal, *TSP Speech Database*, vol. 1. Montreal, QC, Canada: McGill Univ., 2002, pp. 2–09.

[54] J. S. Garofolo, *Timit Acoustic Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.

[55] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Feb. 2011.

[56] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, document Rec. ITU-T P. 862, I.-T. Recommendation, 2001.

[57] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.

[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015. [Online]. Available: https://openreview.net/forum?id=8gmWwjFyLj&fbclid=IwAR1IDmu255FK217fNm1SLg-f5H8YYBzFw53puqJbG1luF_2OChFXEe322qU

[59] S. Venkataramani, C. Subakan, and P. Smaragdis, "Neural network alternatives toconvolutive audio models for source separation," in *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Sep. 2017, pp. 1–6.

[60] A. Gang, P. Biyani, and A. Soni, "Towards automated single channel source separation using neural networks," in *Proc. Interspeech*, Sep. 2018, pp. 3494–3498.

[61] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 1562–1566.

[62] Y. Sun, L. Zhu, J. A. Chambers, and S. M. Naqvi, "Monaural source separation based on adaptive discriminative criterion in neural networks," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, Aug. 2017, pp. 1–5.

[63] S. Qin, T. Jiang, S. Wu, N. Wang, and X. Zhao, "Graph convolution-based deep clustering for speech separation," *IEEE Access*, vol. 8, pp. 82571–82580, 2020.

[64] Y. Wang, J. Du, L.-R. Dai, and C.-H. Lee, "A gender mixture detection approach to unsupervised single-channel speech separation based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 7, pp. 1535–1546, Jul. 2017.

[65] Y. Liu and D. Wang, "A casa approach to deep learning based speaker-independent co-channel speech separation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5399–5403.

[66] S. Vassilvitskii and D. Arthur, "K-means++: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2006, pp. 1027–1035.

[67] A. Feizollah, N. B. Anuar, R. Salleh, and F. Amalina, "Comparative study of k-means and mini batch k-means clustering algorithms in Android malware detection using network traffic analysis," in *Proc. Int. Symp. Biometrics Secur. Technol. (ISBAST)*, Aug. 2014, pp. 193–197.

[68] K. Hornik, I. Feinerer, M. Kober, and C. Buchta, "Spherical k-means clustering," *J. Stat. Softw.*, vol. 50, no. 10, pp. 1–22, 2012.

[69] S. K. Roy, A. Nicolson, and K. K. Paliwal, "DeepLPC-MHANet: Multi-head self-attention for augmented Kalman filter-based speech enhancement," *IEEE Access*, vol. 9, pp. 70516–70530, 2021.

[70] K. Paliwal, K. Wójcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech Commun.*, vol. 52, no. 5, pp. 450–475, May 2010.

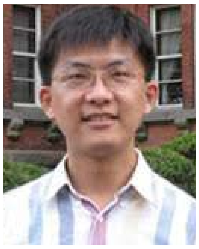
HA MINH TAN received the M.S. degree in electrical and telecommunication engineering from the University of Science, National University HCMC, Vietnam, in 2011. He was a member of the Research Group at the Laboratory of Digital Control and Systems Engineering, National University HCMC. He is currently pursuing the Ph.D. degree in computer science and information engineering with the National Central University, Taoyuan, Taiwan. His current research interests include machine learning, deep learning, blind source separation, computer vision, and bioinformatics.

KAI-WEN LIANG received the M.S. degree in communication engineering from the National Central University, Taoyuan, Taiwan, in 2003, where she is currently pursuing the Ph.D. degree in computer science and information engineering. Her current research interests include signal processing, blind source separation, speech enhancement, machine learning, and deep neural networks.

YUAN-SHAN LEE (Member, IEEE) received the Ph.D. degree in computer science and information engineering from the National Central University, Taoyuan, Taiwan, in 2017. He develops methods for analyzing, extracting, recognizing, and retrieving information from multimedia signals, with special emphasis on speech, image, and music/audio signals. His research interests include source separation, speech enhancement, object recognition, music emotion recognition, sound event recognition, deep neural networks, and Bayesian learning.

CHUNG-TING LI received the M.S. degree in statistics from The National Yang Ming Chiao Tung University Institute, in 2004. He is currently pursuing the Ph.D. degree in computer science and information engineering with the National Central University, Taoyuan, Taiwan. His current research interests include multimedia processing and machine learning.

YUNG-HUI LI (Member, IEEE) received the B.S. degree from National Taiwan University, in 1995, the M.S. degree from the University of Pennsylvania, in 1998, and the Ph.D. degree from the School of Computer Science, Language Technology Institute, Carnegie Mellon University, in 2010. He was a Tenured Faculty Member at National Central University, Taoyuan, Taiwan. He is the Founding Director of the AI Research Center, Hon Hai Research Institute, which is the highest research organization in Foxconn. He is the author of more than 70 conference and journal papers and has written five book chapters. His current research interests include AI, deep learning, machine learning, computer vision, and biometric recognition.



JIA-CHING WANG (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the National Cheng Kung University, Taiwan, in 2002.

He was an Honorary Fellow at the Department of Electrical and Computer Engineering, University of Wisconsin–Madison, in 2008 and 2009. He is currently a Professor with the Department of Computer Science and Information Engineering, National Central University, Taiwan. His research interests include signal processing, deep learning, machine learning, and VLSI architecture design. He is an Honorary Member of Phi Tau Phi Scholastic Honor Society and a member of ACM and IEICE.