Graph Convolution-Based Deep Clustering for Speech Separation

SHAN QIN¹, (Student Member, IEEE), TING JIANG¹, (Member, IEEE), SHENG WU¹, (Member, IEEE), NING WANG², (Graduate Student Member, IEEE), AND XINRAN ZHAO¹

¹Key Laboratory of Universal Wireless Communication, Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China
²Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA 22030, USA

Corresponding author: Shan Qin (qinshan2016@bupt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671075 and Grant 61631003, and in part by the short-term overseas study and exchange project of Ph.D. postgraduates from the Beijing University of Posts and Telecommunications.

ABSTRACT Deep clustering is a promising technique for speech separation that is crucial to speech communication, acoustic target detection, acoustic enhancement and speech recognition. In the study of monophonic speech separation, the problem is that the decrease in separation and generalization performance of the model in the case of reducing the variety of the training data set. In this paper, we propose a comprehensive deep clustering framework that construction the structural speech data based on GCN, named graph deep clustering (GDC) to further improve the separation performance of the separation model. In particular, embedding features are transformed into graph-structured data, and the speech separation mask is achieved by clustering these graph-structured data. Graph structural information aggregates nodes within a class, which makes feature representations conducive to clustering. Experimental results demonstrate that the proposed scheme can improve the clustering performance. The SDR of the separated speech is improved by about 1.2 dB, and the clustering accuracy is improved by 15%. We also use the perceptually motivated objective measures for the evaluation of audio source separation to score the speech quality. The target speech quality and the overall perceptual score are improved by 10.7% compared with other speech separation algorithms.

INDEX TERMS Construction of graph-structured data, deep clustering, graph convolutional filter, speech separation.

I. INTRODUCTION

In the coming intelligent era, human-computer voice interaction technology has been widely concerned. In recent years, how to make human-computer voice interaction as convenient and efficient as human to human communication has become an important topic. Speech separation is essential for the human-computer voice interaction, as its performance has a significant impact on many speech intelligent applications such as acoustic target detection, acoustic enhancement, speech recognition [1], [2]. Therefore, it is of great significance to improve the quality of speech separation.

In recent years, data-driven methods relying on deep neural networks have yielded dramatic improvements in performance compared with the previously used methods, e.g., model-based approaches and matrix factorization [3], [4]. In particular, speaker-independent speech separation has been addressed by approaches like deep clustering, which is the current state-of-the-art [5], [6]. In the deep clustering model, speech separation is regarded as a partition-based segmentation problem [7], which can solve the permutation problem and the dimension mismatch problem. However, the deep clustering algorithm optimizes the similarity between the embedded features and the category labels to extract the distinguishing expression of features, but has no constraint on the in-class distance. Therefore, the problem of label confusion is prone to occur at the category boundary [8], [9]. This may result in insufficient differentiation features limiting the improvement the quality of the separated speech. Some improvement programs are proposed to further improve the quality of the separated speech. To improve the model performance, Chimera++ network adds the loss constraint of the quality of the separated speech in loss function to add the available information [10]. Some method considers modifying the training objective so that it matches that of K-means algorithm used for inference [11]. They use the ratio of the...
in-class variance to the total variance to optimize the model parameters. Other methods have been proposed to whiten the embedded features to improve the separation performance, which would be the Chi-Squared distance between the embedded features and the category labels if the embedded features were a partition. The strategy of the existing improved algorithms is to increase the training difficulty of the networks, which may expand the distance of features belonging to the different sources. However, as a classification problem, the above algorithms don’t consider shortening the in-class distance of features to improve the feature expression ability, which can improve the generalization performance of the speech separation model.

The discrimination of the feature representation plays an important role to improve the quality of separated speech. High discrimination of the feature representations is significant to determine the clustering performance, and separation masks are estimated according to the clustering labels. On the other hand, shortening the in-class distance can effectively improve the differentiation of feature representation. Therefore, we focus on increasing the available information for networks to shorten the in-class distance of features to learn the clustering-friendly features representations. Furthermore, adding long-term memory between time-frequency points can effectively shorten the in-class distance. The in-class distance of features reflects the correlation between the features of the same source. The correlation of features belonging to the same source is not only related to the short-term memory of time-frequency bins, but also to long-term memory of time-frequency bins. However, there are still sequential paths from the past cell to the current cell in order in Long Short Term Memory Network (LSTM), so the memory ability of the network is limited by the gradient disappearance [12], [13]. It means that long-distance correlation of speech extracted only by the LSTM network is limited, as lacking in long-term memory of time-frequency bins may lead to the large in-class distance in the embedded space.

In this paper, we propose a graph deep clustering (GDC) based model to effectively shorten the in-class distance, which can improve speech quality and intelligibility. Specifically, embedding features are transformed into graph-structured data according to the Euclidean distance between features. By integrating the neighbor information of the graph data, long-term dependence between speech features is added to shorten the in-class distance of features. To implement the GDC, a Bi-directional long short term memory (Bi-LSTM) deep clustering model is exploited, where graph convolutional filter (GCF) is integrated into Bi-LSTM layers. The main contributions of this paper can be concluded as follows:

- A graph deep clustering speech separation model is proposed. In this model, the structure information is fused in the embedding features through the graph convolution operation, which shortens the in-class distance and improves the clustering effect. Therefore, the time-frequency bins are closed to each other within class and far away from each other with different sound sources, which improves the performance of the speech separation model.
- A scheme is proposed to transform the non-graphical structure speech data into graph structure data. As speech is non graph structured data, it is necessary to construct the connection of features. According to the similarity of embedding features, we construct the adjacency matrix to establish the long-term correlation of speech. We also analyze the influence of the sparsity of the adjacency matrix on the model performance.

The rest of this paper is organized as follows. Section 2 reviews the background and related work and Section 3 introduces the problem setup. Sections 4 presents our proposed model and an efficient model parameter optimization algorithm, respectively. The experiments on TIMIT data sets are presented in Section 5, and the paper concludes with a summary of the research in Section 6.

II. RELATED WORKS

A. SPEECH SEPARATION BASED ON DEEP CLUSTERING

Deep clustering is a promising technology to the problem of label permutation in speech signal separation [7]. The technique can project each bin of the mixture magnitude spectrogram to a high-dimensional embedding space which is more discriminative for partitioning. Deep clustering approaches have been proposed and showed success in separation tasks [5], [14]. Many techniques reported consisted of multiple stages separately optimized under different criteria, such as signal representation and embedding [15]. Some embedding clustering methods add phase information on multi-channel, and other research concerns the low-delay of deep clustering approaches [16], [17]. Orthogonal deep clustering improves the separation performance of the model by adding an orthogonal constraint penalty term of the objective function to reduce the correlation between the embedded expression [18]. End-to-end approaches have become popular later on, in which all different stages with different functions were jointly trained. K-means algorithm is added to the training layer of the network to make it an end-to-end training based on the deep clustering [19]. Deep attractor networks provide a variety of central point decision strategies to improve separation performance [20]. Some literature have adjusted the clustering algorithm and used different clustering algorithms to improve the performance of the speech separation model [14], [21].

However, none of the above methods takes into account the long-distance correlation of samples to improve the accuracy of clustering. This paper proposes a novel speech separation method exploiting the connection dependency graph of feature, which shortens the in-class distance by establishing long-term correlation and improve the quality of separated speech.

B. GRAPH CONVOLUTIONAL NETWORKS

Graph neural networks (GNNs) are a burgeoning technology to process the structural information of signals [22]. Kipf and
Welling first presented a graph convolutional network (GCN) which proved to be an effective method for extracting meaningful statistical features in graph classification [23]. The essence of graph convolution networks is a Laplacian smoothing low-pass filter, which makes vertex features in the same cluster similar [24]. GCN was also explored for several NLP tasks such as text classification and machine translation, where GCNs were used to encode the syntactic structure of sentences [25]–[27]. However, few works have introduced the GNN technique into the study of monophonic speech separation.

III. PROBLEM SETUP

In clustering-based Time-Frequency masking methods with deep neural networks of embedding generation, the mixture spectrogram \( X \in \mathbb{R}^{T \times F} \) is mapped to a \( D \)-dimensional embedding space

\[
V = H(X)
\]  

(1)

where \( V \in \mathbb{R}^{D \times TF} \) is the embedding for the time-frequency bins, and \( H(\cdot) \) is the mapping function defined by neural networks. The main idea during the training phase is that \( V \) should be clustered into \( C \) classes, each indicating an active source of the mixed speech. This process can be represented as

\[
L = \min \left\| V V^T - Y Y^T \right\|_F^2.
\]  

(2)

The source assignment probabilities \( M = F(V) \in \mathbb{R}^{C \times TF} \) is regarded as the estimated time-frequency mask matrix. Clustering objective is typically aggregated according to a selected type of clustering algorithm \( F(\cdot) \), such that \( M \) matches closely to the target time-frequency masks \( Y \in \mathbb{R}^{C \times TF} \). For hard clustering, ideal binary mask (IBM) \( Y \in \{0, 1\}^{C \times TF} \) is typically used as the target oracle mask [28], while for soft clustering various other masks can be chosen [29]. The separated spectrogram \( \hat{X} \) is obtained by

\[
\hat{X} = X \odot M.
\]  

(3)

As shown by (3), it is important to obtain a high-precision separation mask for improving the quality of speech separation. The differentiation of features directly determines the clustering results. Therefore, it is an essential measure to improve the quality of speech separation to improve the discrimination of feature expression. In addition, for an unseen mixed speech, the model estimates the separated speech by calculating the point product of the mixed speech and the training features in the current embedded space. If we only expand the distance of classes but not shorten the in-class distance, the feature representation of unseen mixed speech would have a weak distinguishability, which will separate speech ineffectively and limit the generalization performance of the model. Therefore, we pay attention to shortening the in-class distance to improve the generalization performance of the model.

IV. PROPOSED MODEL

We propose a graph deep clustering (GDC) model based on a graph convolution filter (GCF). We also present an algorithm to construct the graph structure of speech. The long-term dependence of features is considered to shorten the in-class distance by integrating the neighbor information of the features, which improves the generalization performance of the model.

In this paper, we use the dense association structure of graph information to increase the correlation between samples. Figure 1 is the schematic diagram of the GDC model. It divides into two parts. The first part is the structural representation of embedding features. The second part is speech clustering with a graph convolution filter. The speech data is non-graph structured data. First, it is required to construct an adjacency matrix to transform speech data into graph-structured data. Here, we take the distance of the embedded feature as the connection information of the cell to construct the adjacency matrix of the speech data. Then, the structure features are aggregated by graph convolution. Finally, we cluster the new feature representation to estimate the separation mask of mixed speech. Because the convolution operation is carried out after the model training,
the computational complexity of feature learning does not increase in the training stage. However, as an added module of the whole system, GCF increases the calculation of the separation model.

A. GRAPH-STRUCTURED EMBEDDING REPRESENTATION

The GDC model consisting of structured embedding representation and speech signals clustering is shown in Fig. 1. The structured embedding representation is composed of two parts: 1) The network structure and the improved objective function. 2) The method of constructing structured information on unstructured data.

1) IMPROVED OBJECTION FUNCTION AND NETWORK

To alleviate the decrease of discrimination caused by redundant information of features, we add an over-contracted orthogonal constraint as a regular term, reducing the redundancy of the embedding. We use the over-contracted method to guarantee the orthogonal properties of the embedding like in [18]. We add a hyper parameter $\lambda$ to the unit matrix to achieve fine-tuning. An over-contracted unit diagonal matrix is imposed as a regular term of the objective function

$$L_R = \|V^T V - \lambda I_D\|_F^2.$$  

(4)

Thus, the objective function for GDC is

$$L(V) = \|V^T V\|_F^2 - 2 \|V^T Y\|_F^2 + \|Y^T Y\|_F^2$$

$$+ \|V^T V\|_F^2 - 2\alpha \|V\|_F^2 + \lambda^2 \|I_D\|_F$$  

(5)

where $I_D$ is a $D$-dimensional identity matrix, $\lambda = 1.1$ is a hyperparameter based on experimental experience.

2) CONSTRUCTION OF STRUCTURED DATA

Due to the raw speech is an unstructured data in Euclidean space, we need to convert it into the structured data in non-Euclidean space before processing data by graph filter [30], [31]. For this purpose, we introduce a method to transform the non-graphical structured data onto graph structure data.

Given a graph structure $G = (X, E, A)$ with $V$ the set of nodes ($N = |X|$) and $E \subseteq X \times X$ the set of edges between any two nodes. Let square matrix $A$ represents the adjacency matrix of a graph whose entries $\{a_{ij}\}_{1 \leq j \leq N}$ are the edge weight function such that each edge $e_{ij} \in E$ has a weight $a_{ij}$. We call $A$ the symmetric adjacency matrix ($a_{ij} = 1$ if there is an edge between vertex $x_i$ and vertex $x_j$ ) if we assume that graph $G$ is undirected. An adjacency matrix $A$ is used to represent the connection between nodes in the graph. Here, we construct the connections to speech data based on distance similarity of embedded feature representations of frequency bins.

Let high-dimensional embedding representation of bins $X = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_N) \in \mathbb{R}^{N \times D}$ be nodes with $D$-dimensional features and $\bar{x}_i = [x_{i1}, \cdots, x_{iD}]^T$ be the $ith$ node. The distance $d_{ij}$ between nodes $i$ and $j$ built on a mapping function is

the weight $a_{ij}$ of the edges $e_{ij}$, which is used to construct the adjacency matrix $A$ of embedding shown in (8). The closer the weight is to 1, the higher the similarity between nodes, and vice versa. The distance between each node is defined as in (6)

$$d_{ij} = \frac{\sqrt{(\bar{x}_i - \bar{x}_j)^2}}{2}.$$  

(6)

$$P = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1N} \\ d_{21} & 0 & \cdots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \cdots & 0 \end{bmatrix},$$  

(7)

$$A = Q(P)$$  

(8)

where $P \in \mathbb{R}^{N \times N}$ indicates the distance matrix. $Q(\cdot)$ denotes a mapping function from distance to weights of edges, where it can be represented as

$$Q(P)_{\text{line}} = 1 - \alpha P$$  

(9)

$$Q(P)_{\text{exp}} = e^{-\beta P}$$  

(10)

where $\alpha$ and $\beta$ are attenuation factors. The hyper-parameters $\beta$ could be set according to the data in the experiments, which shows different degrees of attenuation. $\alpha = \max P$ is to guarantee the connection weight of data with the largest distance is 0.

In this paper, we propose two mapping functions to obtain the adjacency weight matrix of feature distance: linear mapping and nonlinear mapping. The formula for linear mapping is shown in (9). The hyperparameter $\alpha$ constraints that the features with the nearest distance have the biggest adjacency weight, and the features with the farthest distance have the smallest adjacency weight. The formula for non-linear mapping is shown in (10). The hyperparameter $\beta$ is an attenuation factor set by experiments. The larger the $\beta$, the faster the attenuation speed. At the same time, it also ensures that the adjacency weight of features with a small distance is greater than the adjacency weight with a long distance.

In Fig. 2, it is shown that different mapping functions have different attenuation for distance. As the distance between the feature vectors increases, the corresponding adjacency
weight value decreases. Vice versa. The exponential function decays faster than the linear function, and the larger the attenuation factor, the faster the decay. To reduce the interference of nodes with low similarity to the current node, we perform sparseness processing on the adjacency matrix. The similarity threshold is determined by the sparsity, i.e., weights greater than the threshold are set to 1 and weights less than the threshold are set to 0. The sparsity is referred to the average degree of nodes in the adjacency matrix in this paper.

An example of the constructed adjacency matrix with different sparsity are given in Fig. 3(a)-3(c). In these examples, a frame of speech data with 129 bins is randomly selected for observation. The similarity of the light areas is higher than that of the dark ones, and we can see that each node has a high similarity with its neighbors. The similarity at the diagonal is the highest. Fig. 3(a) is fully connected adjacency matrix and the sparsity of Fig. 3(c) is the lowest. We also observe the adjacency matrix of the different frames with sparsity = 0.25 shown in Fig. 4(a)-4(c), which demonstrate that the connection between nodes is different from different frames. The connection situation in Fig. 4(a) means that most of the frequency bins of a speech frame may belong to the same sound source for the high similarity with high connection weight. The frequency bins of one speech frame may belong to different sound sources because the similarity of the nodes is low in Fig. 4(b) and Fig. 4(c). That means the constructed adjacency matrix can reflect the similarity between frequency bins. The visualization of the adjacency matrix also shows that there is a correlation between the long-distance frequency bins, which means that the dense connection information exists indeed. Therefore, we can further improve the distinguishing performance of the embedding based on the graph convolution filter constructed by this speech adjacency matrix, and obtain embedding with structural information that is conducive to estimate the separation masks.

B. GRAPH CONVOLUTIONAL FILTER-BASED CLUSTERING

After obtaining the adjacency matrix of the speech data, the structural information can be integrated using the graph filter. Based on graph theory, for the speech separation using GDC, the original GCN model needs to be transformed into GCF, which is to integrate features of nodes to obtain the new feature representations without training. The definition of GCF is provided with Definition 1.

**Definition 1:** GCF is defined as a single weight matrix with one layer and deals with varying node degrees through an appropriate normalization of the adjacency matrix $A$ that improves scalability and classification performance in large-scale networks. The mathematical expressions of GCF $G(\cdot)$ are shown in

$$G(V) = \hat{V} = \hat{A}V$$

where $\hat{A} = \hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}$, $\hat{A} = A + I_N$, $\hat{D} = \sum_j \hat{a}_{jj}$, $\hat{a}_{ij} \in \hat{A}$. $\hat{A}$ indicates a normalization of the adjacency matrix. $\hat{V} \in \mathbb{R}^{N \times TF}$ is the result of embedding representations with structural information processed by GCF.
The node representation contains the information of its first-order neighbors. Structured information of data is integrated into new expressions. The nodes with high similarity will be more clustered, while the nodes that do not belong to the same cluster will be far away from each other. The principal component analysis of features of nodes without GCF are shown in Fig. 5(a) and features of nodes with GCF are shown in Fig. 5(b). Here, we use the fully connected adjacency matrix presented in Section IV-A. In Fig. 5, the distance of nodes can be interpreted as the similarity of embedding. Comparing the distinguishing ability of the two figures at the boundary of categories in the dotted frame, it shows that the distances of different categories in Fig. 5(b) are farther than in Fig. 5(a). That means the new features representations of Fig. 5(b) contains the neighbors information, which helps for clustering.

There is still a long-term memory problem in the processing of sequential signals with LSTM structure. It is not sufficiently to extract the long-term dependent information contained in the time-frequency domain only relying on the LSTM. Therefore, we consider the dense association structure of graph structural data. Each frame of speech is regarded as a graph and the adjacent information is added to shorten the in-class distance of features.

The DC model maps the features of all the participants in the training dataset to a high-dimensional space. Each speaker has a different range of feature representations, so the mixed speech signals can be separated. However, the objective function of DC only expands the distance of distinct classes but not constraining the distance within the same class. This will lead to the confusion of data labels on the boundaries of different categories in the clustering stage, which will reduce the accuracy of clustering and the quality of separated speech.

The new feature representation shortens the distance within the same class, which aggregates the features within the class and increases the accuracy of clustering at the boundary. The new feature representation also improves the generalization performance of the separation model. This is because the expression of the unseen mixed speech is the sum of the feature vectors of all speakers in embedding space. Therefore, the aggregated feature representation is more discriminative for the unseen speakers.

V. EXPERIMENTS

In this section, the performance of the GDC model is evaluated using TIMIT data sets. The experimental setup is introduced, the performance of the proposed model in terms of accuracy and separated speech is evaluated, and the effectiveness of the feature representation on the GDC model is elaborated.

A. EXPERIMENT SETUP

The proposed model uses a network structure based on deep clustering of speech separation [7]. The 4 layers Bi-LSTM network is used in the embedding layer and the local-k means algorithm is applied in clustering. To ensure the fairness of the experiments, the network structures of all the comparison models use 4 layers Bi-LSTM network. Each Bi-LSTM layer has 300 hidden cells and the embedding dimension was set to 20. Stochastic gradient descent with momentum 0.9 and fixed learning rate $10^{-4}$ was used for training. In each updating step, a Gaussian noise with zero mean and 0.7 variance was added to the weight. The total number of epochs was set to be 150. We evaluate the effectiveness of speech separation based on TIMIT data sets on a speech separation task. The data sets, evaluation method, and comparison methods are elaborated in turn.

1) DATA SETS

The TIMIT data sets use as the speech data, which contains 630 speakers with 10 sentences of each one. Here, 30 male voices and 30 female voices were selected and randomly mixed according to different speakers. Verified data sets are mixed with 10 males and 10 females in the same way, and the test database is generated by selecting utterances from 10 unseen people in the training phase. When generating the mixed speech, for the sake of sample imbalance problem, we did some equalization processes when mixing speech.
of different speakers and different genders. So the same number of mixed speeches of male-male, female-female, and male-female combinations is obtained as the training dataset.

All data were down-sampled to 8kHz before processing to reduce computational and memory costs. The input $X$ was the log spectral magnitudes of the mixed speech with $32\text{ms}$ Hanning window length and $8\text{ms}$ window shift. To ensure the local coherency, a mixture is separately processed in half overlapping segments of 100 frames, roughly the length of one word of speech, to output $\hat{V}$ based on the proposed model.

2) COMPARISON METHODS

- **Deep clustering:** The original model is represented as DC. DC w/o tanh represents the model without the activation function of the output layer based on DC. The DC model employs $K - \text{means}$ clustering to determine the cluster center and get estimated separation masks in the test phase.

- **Anchored deep attractor network:** The anchored deep attractor network is represented as ADAN. The ADAN model calculates the anchors to determine the location of the clustering center.

- **Graph deep clustering:** The GDC model is the proposed model. We continue to use $K - \text{means}$ clustering in the test phase to estimate the separation masks.

- **Baseline:** The separated speech obtained from the Ideal ratio mask (IRM) serves as the reference baseline.

3) EVALUATION METRIC

We validate the performance of the proposed model from two aspects: speech separation quality evaluation index and clustering accuracy. The PEASS software provides a set of perceptually motivated objective measures for the evaluation of audio source separation. Overall perceptual score (OPS) is an overall evaluation of the separated speech, whose score is a percentage system. $Q_{\text{target}}$ is related to the salience of the distortion of the target source of the source estimate. If the value is close to 1, the distortion is small. Moreover, the signal distortion ratio (SDR) is used as a speech separation index.

**TABLE 1.** Quality evaluation metrics of separated speech on different models.

| Models | $Q_{\text{target}}$ | OPS | SDR (dB) |
|--------|---------------------|-----|---------|
| DC     | 0.787               | 85  | 7.9     |
| DC w/o tanh | 0.799       | 89  | 8.3     |
| ADAN   | 0.797               | 92  | 8.6     |
| GDC$_{\text{relu}}$ | 0.802     | 98  | 9.1     |
| IRM    | 1                   | 100 | -       |

**TABLE 2.** Quality evaluation metrics of separated speech with different activation function on the GDC model.

| Models | $Q_{\text{target}}$ | OPS | SDR (dB) |
|--------|---------------------|-----|---------|
| GDC$_{\text{tanh}}$ | 0.789        | 87  | 8.7     |
| GDC$_{\text{non}}$ | 0.865       | 93  | 8.3     |
| GDC$_{\text{elu}}$ | 0.812       | 94  | 8.9     |
| GDC$_{\text{relu}}$ | 0.802      | 98  | 9.1     |

**TABLE 3.** Quality evaluation metrics of separated speech with different activation function on different models.

| Metrics | Models | Activation function | $Q_{\text{target}}$ | OPS | SDR |
|---------|--------|---------------------|---------------------|-----|-----|
|         | GDC    | tanh                | 0.85                 | 0.887 | 0.855 | 0.921 |
|         | DC     | non                 | 0.789               | 0.865 | 0.812 | 0.802 |
|         | GDC    | relu                | 0.91                 | 0.91  | 0.93  | 0.93  |
|         | DC     | elu                 | 0.85                 | 0.89  | 0.79  | 0.85  |
|         | GDC    | tanh                | 0.85                 | 0.89  | 0.92  | 0.102 |
|         | DC     | non                 | 0.79                 | 0.83  | 0.83  | 0.845 |

**B. PERFORMANCE**

1) EFFICIENCY EVALUATION

a: COMPARISON OF QUALITY OF SEPARATED SPEECH

On the small data training set, the validity of the GDC model is verified by comparing with the existing algorithm for speech quality index.

As shown in Table 1, after removing the activity function of the last layer, the effect of the system is improved obviously. The separation performance of the GDC model with $\text{relu}(-)$ is closer to the results of IRM than DC models. This also reveals that the GDC model can effectively improve the clarity and intelligibility of separated speech.

Table 2 shows the effect on the performance of the model when different activation functions are used in the output layer of the GDC model. It can be seen from the quality evaluation metrics of the separated speech that the performance of the $\text{relu}(-)$ function is the best, and the $\text{elu}(-)$ and $\text{non}$ activation function are suboptimal. When using the $\tanh(-)$ as activation function of the output layer, the model performance is significantly lower than the above three methods. That is to say, adjusting the activation function of the output layer enables the network to effectively train the unseen training data sets and improve the speech separation performance of the model.

We compare the quality evaluation metrics of separated speech with different activation function on different models. From the comparison and analysis of the metrics in the Table 3, it shows that compared with the DC model, the SDR of GDC model is increased by about 10%, and the OPS is increased about 10.7%, and the overall performance of the GDC model is optimized.

And it also shows that the different activation functions played a positive role in improving model performance. For the models trained by the $\text{relu}(-)$ series of functions, the quality of speech separation is higher than that of $\tanh(-)$ in both models. Further analysis the results show that even though the DC model with $\text{elu}(-)$ outperforms than the other DC models, it still has a slightly worse than the GDC model on SDR performance. Therefore, the new feature representation with the graph structure information works well in improving the model performance.

Visualizations of the quality evaluation metrics $Q_{\text{target}}$, OPS and SDR are shown in Fig. 6. The figure reveals that...
when the number of speakers is 60 and 80, the performance of the GDC model is significantly better than the DC model. As the number of speakers increases, both models have an upward trend. But the performance of the GDC model is relatively stable. The DC model is clearly on the rise. In other words, the generalization performance of the DC model is significantly reduced on a small training set with a small number of speakers. It also shows that the generalization performance of our proposed model is superior to the DC model. And the model performance is stable on a small training set.

From the above experimental analysis, it can be known that the feature representation combined with graph structure information is more conducive to the estimation of the speech separation mask. The graph structure information shortens the intra-class distance and increases the aggregation of speech feature representations of the same source. The new features are clustering-friendly to obtain the more accurate estimated masks, and the quality of the separated speech are also improved naturally.

b: COMPARISON OF CLUSTERING ACCURACY
The accuracy and the precision of clustering for embedding labels in the test stage are calculated according to the IRM of pure speech provided by training data.

In our experiments, we calculated the average accuracy per 100 frames as experimental data for comparative analysis. Every 100 frames have 12900 frequency bins corresponding to a clustering result. For example, if a sentence has \( n \) frames, the clustering results are averaged and the formula is expressed as follows

\[
P = \frac{1}{N} \sum_{i=1}^{N} p_i \tag{12}
\]

where \( N = \frac{n}{100} \) and \( n \) is the total frame length of the test speech. \( p_i \) is the clustering accuracy or accuracy of each 100 fragment. It is shown in Fig.4 that fuse structured information will be helpful to improve the accuracy of the label assignment. In Fig. 7, the clustering accuracy and precision of the GDCrelu model are higher than those of the original model, which shows that the embedding after GCF is more conducive to clustering. The accuracy of clustering is improved by 15% and the score of speech perception evaluation is improved by 9.3%. The validity of the model is further verified by combining the evaluation score of separated speech quality.

The low accuracy and precision of clustering lead to the label permutation problems. Therefore, the original DC model on small training data sets causes the label confusion problem and gets the label-cross separated speech. The GDC with \( \text{relu}(\cdot) \) series function models on small training data sets increase the accuracy and precision of clustering, which overcomes the label ambiguity problem and obtains the complete separated speech instead of label-cross separated speech. Therefore, small training data sets are effectively trained in the GDC with the \( \text{relu}(\cdot) \) series function models.

Comparison of the convergence performance of different activation functions: We evaluate the effectiveness of different activation function of the output layer in GDC models.

In Fig. 8 depicts the values of the loss function of GDC models with different activation functions during the training phase. The distribution of the loss function reflects the
convergence performance of the model. We compared four cases: the \text{relu}(\cdot) function, the \text{elu}(\cdot) function, and the \text{tanh}(\cdot) and \text{non} activation function. As shown in the figure, the convergence rates and the loss values of the \text{relu}(\cdot) series function are faster and lower than the other two cases, which indicates that the training with the \text{relu}(\cdot) series function is more effective. The convergence of \text{relu}(\cdot) is the best of the four cases.

2) HYPERPARAMETER SENSITIVITY TEST

Comparison of different sparsity of adjacency matrix: We discuss the influence of different sparsity of the adjacency matrix on speech separation. Fig. 9 and Fig. 10 indicate that select appropriate neighbors for feature fusion can improve the separation performance of the model. Fully connected adjacency matrices may introduce interference information and weaken the separation performance. When the sparsity is 0.3, the performance is better than other cases.

VI. CONCLUSION

The proposed model adds the long-term dependence between time-frequency bins by the graph convolution operation to shorten the in-class distance effectively, and the generalization performance of the model is improved. The GDC model proposed in this paper has two unique characteristics: (1) the structural information of the data is added into the deep clustering model. GCF is used to integrate the neighborhood information of nodes to shorten the in-class distance. (2) The model works well on the unseen test sets for improving the generalization performance. And adding the structure information also reduces the model dependence on the specific data sets.

REFERENCES

[1] R. Zazo, P. S. Nidadavolu, N. Chen, J. Gonzalez-Rodriguez, and N. Dehak, “Age estimation in short speech utterances based on LSTM recurrent neural networks,” IEEE Access, vol. 6, pp. 22524–22530, 2018.
[2] W. Jiang, F. Wen, and P. Liu, “Robust beamforming for speech recognition using DNN-based time-frequency masks estimation,” IEEE Access, vol. 6, pp. 52385–52392, 2018.
[3] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, “Speaker recognition for multi-speaker conversations using X-vectors,” in Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 5796–5800.
[4] Y. Jia, M. Johnson, W. Macherey, R. J. Weiss, Y. Cao, C.-C. Chiu, N. Ari, S. Laurenzo, and Y. Wu, “Leveraging weakly supervised data to improve end-to-end speech-to-text translation,” in Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 7180–7184.
[5] R. Aihara, T. Hanazawa, Y. Okato, G. Wichern, and J. L. Roux, “Teacher-student deep clustering for low-delay single channel speech separation,” in Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 690–694.
[6] C. Han, Y. Luo, and N. Mesgarani, “Online deep attractor network for real-time single-channel speech separation,” in Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 361–365.
[7] J. R. Hershey, Z. Chen, J. Le 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.
[8] Q. Wang and G. Guo, “Benchmarking deep learning techniques for face recognition,” J. Vis. Commun. Image Represent., vol. 65, Dec. 2019, Art. no. 102663.
[9] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, “CosFace: Large margin cosine loss for deep face recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5265–5274.
[10] 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.
[11] 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.
[12] S. H. I. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 892–810.
[13] M. Coto-Jiménez and J. Goddard-Close, “LSTM deep neural networks postfiltering for enhancing synthetic voices,” Int. J. Pattern Recognit. Artif. Intell., vol. 32, no. 1, Jan. 2018, Art. no. 1860008.
[14] Y. Luo and N. Mesgarani, “Augmented time-frequency mask estimation in cluster-based source separation algorithms,” in Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 710–714.
[15] C. Fan, B. Liu, J. Tao, J. Yi, and Z. Wen, “Discriminative learning for monaural speech separation using deep embedding features,” 2019, arXiv:1907.09884. [Online]. Available: http://arxiv.org/abs/1907.09884
[16] S. Wang, G. Naithani, and T. Virtanen, “Low-latency deep clustering for speech separation,” in Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 76–80.
[17] P. Seetharaman, G. Wichern, J. Le Roux, and B. Pardo, “Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures,” in Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), May 2019, pp. 356–360.
S. Qin et al.: Graph Convolution-Based Deep Clustering for Speech Separation

[18] S. Choe, S.-W. Chung, Y. Ji, and H.-G. Kang, “Orthonormal embedding-based deep clustering for single-channel speech separation,” 2019, arXiv:1901.04690. [Online]. Available: http://arxiv.org/abs/1901.04690

[19] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” 2016, arXiv:1607.02173. [Online]. Available: http://arxiv.org/abs/1607.02173

[20] 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.

[21] S. Ke, R. Hu, G. Li, T. Wu, X. Wang, and Z. Wang, “Multi-speakers speech separation based on modified attractor points estimation and GMM clustering,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2019, pp. 1414–1419.

[22] P. W. Battaglia et al., “Relational inductive biases, deep learning, and graph networks,” 2018, arXiv:1806.01261. [Online]. Available: http://arxiv.org/abs/1806.01261

[23] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” 2016, arXiv:1609.02907. [Online]. Available: http://arxiv.org/abs/1609.02907

[24] Q. Li, X.-M. Wu, H. Liu, X. Zhang, and Z. Guan, “Label efficient semi-supervised learning via graph filtering,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 9582–9591.

[25] R. Wang, H. Zhao, S. Ploux, B.-L. Lu, M. Utiyama, and E. Sumita, “Graph-based bilingual word embedding for statistical machine translation,” ACM Trans. Asian Low-Resource Lang. Inf. Process., vol. 17, no. 4, pp. 1–23, Aug. 2018.

[26] Y. Li, R. Jin, and Y. Luo, “Classifying relations in clinical narratives using segment graph convolutional and recurrent neural networks (SegGCNs),” J. Amer. Med. Inform. Assoc., vol. 26, no. 3, pp. 262–268, Mar. 2019.

[27] R. Angelova and G. Weikum, “Graph-based text classification: Learn from your neighbors,” in Proc. 29th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR), 2006, pp. 485–492.

[28] A. Plinge, M. H. Hennecke, and G. A. Fink, “Reverberation-robust online multi-speaker tracking by using a microphone array and casa processing,” in Proc. Int. Workshop Acoustic Signal Enhancement (IWAENC), Sep. 2012, pp. 1–4.

[29] D. Enami, F. Zhu, K. Yamamoto, and S. Nakagawa, “Soft-clustering technique for training data in age-and gender-independent speech recognition,” in Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf., Dec. 2012, pp. 1–4.

[30] E. van Dellen, L. Douw, J. C. Baayen, J. J. Heimans, S. C. Ponten, W. P. Vandertop, D. N. Velis, C. J. Stamm, and J. C. Reijneveld, “Long-term effects of temporal lobe epilepsy on local neural networks: A graph theoretical analysis of corticography recordings,” PLoS ONE, vol. 4, no. 11, p. e8081, 2009.

[31] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A comprehensive survey on graph neural networks,” 2019, arXiv:1901.00596. [Online]. Available: http://arxiv.org/abs/1901.00596

TING JIANG (Member, IEEE) was born in January 1962. He received the B.S., M.S., and Ph.D. degrees from Yanshan University, in 1982, 1988, and 2003, respectively. He is currently a Professor with the Beijing University of Posts and Telecommunications, Beijing, China. He has presided over and participated in more than ten national natural fund, major special projects and enterprise projects, published more than 50 related research articles in academic journals and conferences at home and abroad, and obtained six authorized patents. His research interests are mainly in wireless broadband interconnection, information theory, theory and application of short-range wireless communication technology, and wireless sensor networks.

SHENG WU (Member, IEEE) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications, Beijing, China, in 2004 and 2007, respectively, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, in 2014. He was a Postdoctoral Researcher with the Tsinghua Space Center, Tsinghua University. He is currently an Associate Professor with the Beijing University of Posts and Telecommunications. He has published over 60 journal and conference papers. He also holds over 30 granted patents. His research interests are mainly in iterative signal processing, satellite communications, and mobile communication networks. He has received the First Prize of Science and Technology Award of the Chinese Institute of Electronics, in 2017, the Silver Medal of the 46th Geneva International Exhibition of Inventions, in 2018, and the Second Prize of National Technological Invention Award of China, in 2018.

NING WANG (Graduate Student Member, IEEE) received the Ph.D. degree in information and communication engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2017. He was an Engineer with Huaxin Post and Telecommunications Consulting Design Company, Ltd., Hangzhou, Zhejiang, China, from 2012 to 2013. He is currently a Postdoctoral Scholar with the Electrical and Computer Engineering Department, George Mason University, Fairfax, VA, USA. His current research interests are in physical layer security, machine learning, device identification, and RF fingerprinting.

SHAN QIN (Student Member, IEEE) received the B.S. and M.S. degrees in electronics and communication engineering from Tianjin Normal University, Tianjin, China, in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree in information and communication engineering with the Beijing University of Posts and Telecommunications, Beijing, China. Her research interests include the signal processing, machine learning, data mining, and knowledge mapping.

XINRAN ZHAO is currently pursuing the bachelor’s degree, majoring in electronic and communication engineering, with the Beijing University of Posts and Telecommunications, mainly studies blind source separation of multiple speakers.

* * *