Multi-Head Self-Attention-Based Deep Clustering for Single-Channel Speech Separation

YANLIANG JIN, CHENJUN TANG, QIANHONG LIU, AND YAN WANG

Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China

Corresponding author: Yanliang Jin (wuhaide@shu.edu.cn)

This work was supported in part by the NSFC, China, under Grant 61771299, in part by the National Key Research and Development Program of China under Grant 2018YFB2101303, in part by the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Shanghai University, under Grant SKLSFO2012-14, in part by the Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, in part by the Shanghai Education Committee, Chinese Academy of Sciences, in part by the Shanghai Science Committee under Grant 12511503303, Grant 14511105602, and Grant 19511102803, in part by the H2020 under Grant 778305, and in part by the Innovate U.K., under Grant 10754.

ABSTRACT Turning attention to a particular speaker when many people talk simultaneously is known as the cocktail party problem. It is still a tough task that remained to be solved especially for single-channel speech separation. Inspired by the physiological phenomenon that humans tend to distinguish some attractive sounds from mixed signals, we propose the multi-head self-attention deep clustering network (ADCNet) for this problem. We creatively combine the widely used deep clustering network with multi-head self-attention mechanism and exploit how the number of heads in multi-head self-attention affects separation performance. We also adopt the density-based canopy K-means algorithm to further improve performance. We trained and evaluated our system using the Wall Street Journal dataset (WSJ0) on two and three talker mixtures. Experimental results show the new approach can achieve a better performance compared with many advanced models.

INDEX TERMS Single-channel speech separation, deep clustering, multi-head self-attention, density-based canopy K-means.

I. INTRODUCTION Cocktail party problem which is of great significance for automatic speech recognition and voiceprint recognition [1] was first proposed by Cherry in [2]. Conventional solutions for single-channel speech separation mainly include computer auditory scene analysis (CASA) [3], [4] and non-negative matrix factorization (NMF) [5]. CASA simulates the processing of sound by human auditory system using certain organizational guidelines and appropriate separation cues. NMF is based on the assumption that audio spectrogram has a low-rank configuration and can be represented with a small number of bases [5]. However, these traditional techniques suffer from similar issues relating to performance with limitations on spectral dynamics or the exploitation of temporal, unknown speakers, and high complexity [6]. Recently, deep learning has been successfully applied to source separation. Deep neural networks (DNNs) were adopted to capture the highly non-linear relationship of speech characteristics between mixed signals and the target speaker [7]–[10]. Deep clustering (DPCL) was creatively exploited for monaural source separation in [11]. The essence of DPCL is to determine source assignment based on a similarity measurement in original spectral space or embedding space. Following DPCL, many innovative approaches have been successfully applied to formulate separation as a regression problem and learn an effective mapping from the mixture to the source time-frequency masks. Typical solutions include permutation invariant training (PIT) [12], utterance-level permutation invariant training (uPIT) [13] and deep attractor network (DANet) [14]. However, the reconstruction of source signal with time-frequency masking is imperfect. A time-domain audio separation network (TasNet) [15] was conducted and achieved good results. Furthermore, a joint audio-visual model for speech separation was put forward in [16] and it outperforms other audio-only models. However, it is notable that in these methods, the input mixed speech is considered as a pure engineering signal and the physiological characteristics of human beings are ignored. For instance, humans can concentrate on one or two sounds...
of interest in a noisy party or other complex auditory environments. It seems easy for humans to recognize a single voice from a signal consisting of multiple sources. According to the research on cortical representation of multi-talker mixed speech by Mesgarani and Chang [17], human auditory system can restore the representation of the speaker of interest while controlling irrelevant competing speech when several speakers talk simultaneously. In other words, human auditory system will assign different weights to different units in the speech sequence. This feature is similar to the attention mechanism which is widely used in natural language processing (NLP) tasks. It was initially proposed in [18] for machine translation. It works as a pooling layer which decides the most discriminative features over sequence and learn the distribution of weights by computing the similarity of elements in sequence. So in our paper, we try to apply the attention mechanism to address the cocktail party problems. We conduct the joint optimization of DPCL model [11] and multi-head self-attention, which can jointly attend to information from different representation subspaces at different positions. The main contributions of this paper are: 1. Human beings can separate target speech from mixed signals easily owing to the speciality of auditory system. We utilize this physiological characteristic for cocktail party problem. 2. We creatively combine the original model with multi-head self-attention mechanism. Multi-head self-attention can capture relevant information on different subspaces by calculating multiple times so that the network can learn more comprehensive information. Besides, how the number of heads in multi-head self-attention mechanism impacts model performance is also explored. 3. K-means is used in DPCL model. However, the value of $K$ should be known at first, and it is not effective in dealing with big data. So we adopt the density-based canopy K-means clustering algorithm, which not only improves the final performance but also makes the system compatible with the condition where the number of speakers is unknown. The organization of this paper is as follows: Section II introduces the original DPCL model and its drawbacks. Section III introduces the proposed methods, including the multi-head self-attention mechanism and density-based canopy K-means. Section IV presents the experimental setting and results using the Wall Street Journal dataset (WSJ0) speech corpus. In Section V, experimental conclusions and future work directions are described.

II. DPCL MODEL

DPCL for single-channel speech separation was first put forward by Hershey et al., etc., [11]. The essence of it is to train a neural network to learn the high-dimensional embedding of each time-frequency unit so that the embedding belonging to the same speaker has the minimum distance in the embedding space. The value of complex spectrogram at corresponding time-frequency bin is defined as: $X_t = X_{tf}, i \in (1, \ldots, N)$, where $t$ and $f$ represent the time index and frequency index of the signal respectively. A single time-frequency bin $X_{tf}$ and a frame $X_s$ are mapped to $p_{sf}$ and $P_{is}$ by deep neural network, where $p_{sf} \in \mathbb{R}^{Dx1}, P_{is} \in \mathbb{R}^{DFx1}$ and $|p_s|^2 = 1$. $D$ represents the dimension of embeddings. $Y = \{y_{i,c}\}$ is a one-hot label vector indicating which source dominates the mixture. It maps element $i$ to $c$ categories. When element $i$ belongs to class $c$, $y_{i,c} = 1$, otherwise it is 0. $YY^T$ is a binary affinity matrix representing the clustering results. When element $i$ and element $j$ belong to the same category, $(YY^T)_{ij} = 1$, otherwise it is 0. DPCL defines an objective function to make the estimated affinity matrix as close as possible to the real one. This formula is expressed as:

$$C = ||(PP^T - YY^T)||_F^2 = \sum_{i,j} (|p_i - p_j|^2 - 1) + \sum_{i,j:y_i = y_j} (|p_i - p_j|^2)^2$$

(1)

![FIGURE 1. The architecture of DPCL.](image)

The structure of DPCL is shown in figure 1. The high dimensional embeddings are extracted by bidirectional long short-term memory (Bi-LSTM) network, then K-means is utilized to separate the time-frequency bins. It should be particularly emphasized that the K-means clustering is used only at testing time, which is different from other source separation methods learning a direct mapping from mixed signals to source signals. Bi-LSTM has two parallel long short-term memory (LSTM) layers propagating in two directions. It computes the sequence from the front to the back to get the forward hidden state $h_f$, then from the back to the front to get the reverse hidden state $h_r$. The final hidden state
of the sequence is calculated by merging the forward hidden state and the reverse hidden state of each element according to some merging strategies such as splicing, averaging and summation. However, traditional LSTM unifies the hidden variable-length vector into a fixed-length vector [19]. In this way, speech vectors of different lengths are encoded into feature vectors of fixed length, which results in the loss of intermediate nodes in LSTM, especially in processing long speech. What is more, temporal pooling averages these representations over time to obtain a final encoded embedding. The major drawback of this method is that every element of the sequence must contribute equally in obtaining the representation. So a trainable layer is necessary for assigning a weight over each representation of the sequence. Besides, in the K-means clustering algorithm, the value of K should be known at first which is impractical in real life. Based on above problems, we adopt the multi-head self-attention mechanism, which can learn features of a sequence from different aspects and retain the output of hidden units in LSTM. Besides, the original K-means clustering algorithm is also improved in our model.

III. MODEL

The overview architecture of our proposed model is shown in figure 2. It mainly consists of four parts: pretreatment layer, embedding network, multi-head self-attention layer, improved K-means layer. The last two parts of our proposed model will be presented in detail in the following sections.

A. ATTENTION

Recently, attention mechanisms have become an integral part of models that can model the strength of relevance between representation units. The input of attention mechanism consists of queries \( Q \), keys \( K \), and values \( V \). Its structure is shown in figure 3. \( i \in [1, L] \), \( L \) is the length of input vector. There are three main steps in calculating attention value: (1) Calculate the similarity between query \( Q \) and each key \( K_i \) to get the weight. The commonly used methods for similarity calculation are as follows:

\[
f(Q, K_i) = \begin{cases} Q^T K_i & \text{dot product} \\ Q^T W_a K_i & \text{general} \\ W_a [Q^T; K_i] & \text{concat} \end{cases}
\]

(2) Normalize the weights using the softmax function.

\[
\alpha_i = \text{softmax}(f(Q, K_i)) = \frac{\exp(f(Q, K_i))}{\sum_i \exp(f(Q, K_i))}
\]

(3) The weight and corresponding value are weighted and summed to get the final attention result. Where \( Q, K, V \) denote query matrix, keys matrix and value matrix respectively.

\[
\text{Attention}(Q, K, V) = \sum_{i=1}^{L} \alpha_i \ast V_i
\]
which enhances the importing of information. Reference [25] employs the multi-head attention to highlight the speaker related features learned from context information in frequency and time domain. As shown in figure 4, for given sequence \( R \), the multi-head attention mechanism relies on scaled dot-product attention, which operates on queries, keys, values of dimension \( d_k, d_k, d_v \) with linear projections. By stacking them into matrices \( Q, K, V \), the output of attention is:

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\] (5)

In self-attention, \( Q = K = V \) and they are from the output of the previous layer. To exploit information from different representation subspaces at different positions, multi-head self-attention is further proposed to perform multiple self-attention function \( h \) times to generate queries, keys, values matrices \( Q_i, K_i, V_i \) from \( i = 1, \ldots, h \). So \( h \) also can be considered as the number of heads in multi-head self-attention. The multi-head self-attention is calculated as:

\[
\text{head}_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i) \quad (6)
\]

\[
\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^o \quad (7)
\]

where \( W^o, W^Q_i, W^K_i, W^V_i \) are trainable weight matrices. Since each attention layer captures different modalities, it is expected to improve the performance with an ensemble effect. But there is a potential problem: the network will become more complex and its performance even will deteriorate with the increasing of \( h \). In our following paper, we combine the DPCL model in [11] with multi-head self-attention and explore the most proper value of \( h \) for our proposed model.

B. DENSITY-BASED CANOPY K-MEANS

K-means algorithm is a universally used clustering algorithm based on division concept, it is simple and efficient but the determination of cluster number and the selection of initial clustering centers remain hard [26]. While canopy
clustering is of low accuracy, it does not need to specify the number of K values [27]. So density-based canopy can be employed as pre-processing procedure, then its result is utilized as cluster number and initial clustering center of K-means algorithm [28]. The sample elements m in dataset \( M = \{x_1, x_2, \ldots, x_n\} \) is \( x_m = (x_{m1}, x_{m2}, \ldots, x_{mr}) \). \( n \in [1, N] \), \( N \) is the number of elements in \( M \). The average distance of all elements in \( M \) is defined as:

\[
S = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=i+1}^{N} d(x_i, x_j) \quad (8)
\]

\[
d(x_i, x_j) = \sqrt{\sum_{k=1}^{r} (x_{ik} - x_{jk})^2} \quad (9)
\]

where \( d(x_i, x_j) \) represents the Euclidean distance between \( x_i \) and \( x_j \). As illustrated in (10), \( \rho(i) \) is the number of data whose distance from other samples \( j \) to point \( i \) is less than \( S \). \( f(x) \) is a boolean function whose value is set to 1 when \( x < 0 \) and 0 when \( x > 0 \).

\[
\rho(i) = \sum_{j=1}^{N} f[d(i, j) - S] \quad (10)
\]

\( a(i) \) is the average distance between samples in cluster and it is defined in (11).

\[
a(i) = \frac{2}{\rho(i)(\rho(i) - 1)} \sum_{i=1}^{\rho(i)} \sum_{j=i+1}^{\rho(i)} d(x_i, x_j) \quad (11)
\]

\( O(i) \) stands for the distance between element \( i \) and other element \( j \) with a higher local density. \( O(i) \) is defined in (12). If existing \( \rho(j) > \rho(i) \), \( O(i) \) is defined as \( \min \{d(i, j)\} \), else it is \( \max \{d(i, j)\} \).

\[
O(i) = \begin{cases} 
\min \{d(i, j)\}, & \exists j, \rho(j) > \rho(i) \\
\max \{d(i, j)\}, & \forall j, \rho(j) > \rho(i)
\end{cases} \quad (12)
\]

\( \omega \) is defined as the weight product of \( \rho(i) \), \( a(i) \) and \( O(i) \) in (13).

\[
\omega = \rho(i) \ast \frac{1}{a(i)} \ast O(i) \quad (13)
\]

According to the practical significance of \( \rho(i) \), \( a(i) \) and \( O(i) \), we can conclude: (a) the greater the value \( \rho(i) \), the more the samples gather around the point \( i \) and the tighter the samples in a cluster; (b) The smaller the value \( a(i) \), the larger the value \( 1/a(i) \), the more intensive the cluster will be; (c) The larger the value \( O(i) \), the greater the degree of dissimilarity between the two clusters. So in order to divide the datasets appropriately, \( \omega \) needs to be maximized. At first, the density of every element in the dataset \( M \) is calculated according to (10). The maximum value of \( \rho \) will be the first cluster center \( c_1 \), all samples whose distance to the initial cluster center is less than \( S \) will be added to the new cluster \( C_1 \) and at the same time, they will be removed from the original dataset \( M \). For the rest samples in dataset \( M \), \( \rho(i) \), \( a(i) \) and \( O(i) \) are calculated according to (10)-(11). The second clustering \( c_2 \) center will be decided to maximize the value of \( \omega \). The data meet the conditions will be removed from \( M \) and added to the cluster \( C_2 \), the same as the procedures in \( C_1 \). For the rest elements in \( M \), the distance between them and the points in \( C = \{c_1, c_2\} \) will be calculated. The third clustering center will be determined to maximize \( \omega(i, c_1) \ast \omega(i, c_2) \), where \( i \) is the index of the sample in \( M \). Then elements meeting the condition will be removed from the old dataset and be added to the new cluster. The above process will be repeated until the dataset \( M \) is empty. At last, the original dataset will be divided into \( K \) subsets. The clusters and the value of \( K \) then will be input to the K-means algorithm to be clustered more precisely.

IV. EXPERIMENTS

A. DATASETS

The training and evaluation for our model are based on the Wall Street Journal (WSJ0) dataset. In the original WSJ0 dataset, the training part contains fifty men speakers and fifty-one women speakers while the testing part includes ten men speakers and eight women speakers. Every speaker has 141 or 142 utterances, each of which lasts for about five or six seconds. The bit rate of the voices is 256 kb/s, and the sample rate is 16 KHz. We trained and evaluated our network on two-speaker speech separation problem using WSJ0-2mix dataset [11], which contains 18000 utterances and lasts about 30 hours. The mixtures were generated by randomly selecting speech from different speakers in WSJ0 training set \( si_tr_s \), and mixing them at random signal-to-noise ratios (SNR) between −5 dB and 5 dB. Five hours of evaluation set was generated in the same way, using utterances from 16 unseen speakers from \( si_dt_05 \) and \( si_et_05 \) in the WSJ0 dataset. The WSJ0-3mix dataset was generated using a similar approach but it contains mixtures of speeches from three speakers. For most existing speech separation methods, the identity of speakers should be known at first, while the task for unknown speakers separation is challenging. Besides, since the mixture of different gender speakers has different separation performance, we created known and unknown speakers datasets which include different gender speakers in the WSJ0 test section. Therefore our test dataset includes known speakers set and unknown speakers set, both of which can be divided into two male speakers mixed sets, two female speakers mixed sets, a male speaker and a female speaker mixed set.

B. SETUP

In pretreatment layer, every audio file is downsampled to 8 KHz. The log short-time Fourier spectral magnitudes of mixture speech are extracted by the hamming window with a 32 ms window length and 8 ms window shift. Each utterance is divided by 100 frames, which is roughly the length of one word. To avoid silent speech, some speech bins whose amplitude is 40 dB lower than speaker’s maximum amplitude are removed. 256-point DFT is applied to each frame of Bi-LSTM training to extract 129-dimensional log-
To evaluate the effect of multi-head self-attention layer and normal K-means (*ADCNet) and DPCL with multi-head attention layer and canopy K-means (ADCNet) in WSJ0-2mix.

| Model                      | known speakers set | female+male | all | male+male | unknown speakers set | female+male | all |
|----------------------------|--------------------|-------------|-----|-----------|----------------------|-------------|-----|
| DPCL+K-means[11]           | 3.79               | 4.29        | 9.04| 6.34      | 5.49                 | 3.21        | 8.69| 6.45|
| DPCL+canopy-K-means        | 3.92               | 4.55        | 9.28| 6.71      | 6.68                 | 3.54        | 8.87| 6.70|
| *ADCNet                    | h=1 10.33          | 10.45       | 11.29| 10.94     | 10.21                | 10.27       | 10.86| 10.72|
| ADCNet                     | h=1 10.51          | 10.88       | 11.52| 11.21     | 10.33                | 10.49       | 11.18| 10.94|

SI-SNR is defined as:

\[ SI - SDR := 10 \log_{10} \left( \frac{|S|^2}{|S|^2 + |\hat{S} - \hat{S}|^2} \right) \]  

where \( S \in \mathbb{R}^{1 \times t} \) and \( \hat{S} \in \mathbb{R}^{1 \times t} \) denotes the estimated and original clean source respectively, \( t \) is the length of signals, and \( |S|^2 = \langle S, S \rangle \) denotes the power of the signal. \( e_{\text{interf}} \), \( e_{\text{noise}} \), \( e_{\text{artif}} \) are the interferences, noise and artifacts error terms respectively. \( \hat{S} \) and \( S \) are both normalized to have zero-mean to ensure scale-invariance.

**C. EXPERIMENTAL RESULTS AND ANALYSIS**

SDR results about different gender combinations and overall performance across all combinations on the original DPCL model [11], DPCL with density-based canopy K-means, *ADCNet and ADCNet were recorded in Table 1. SI-SDR, PESQ, SI-SNR results were shown in Table 2, Table 3 and Table 4 respectively.

1) THE EFFECT OF MULTI-HEAD ATTENTION LAYER AND DENSITY-BASED CANOPY K-MEANS

According to Table 1, by comparing the results of “DPCL+K-means” and “DPCL+canopy K-means”, “DPCL+K-means” and “*ADCNet (h = 1, 2, 3, 4)”, it is distinct that both multi-head self-attention mechanism and density-based canopy K-means algorithm can improve performance, while multi-head self-attention is much more effective. By compar-
TABLE 2. SI-SDR (dB) results with different \( h \) for DPCL with multi-head attention layer and normal K-means (*ADCNet) and DPCL with multi-head attention layer and canopy K-means (ADCNet) on unknown speakers in WSJ0-2mix.

| Model | male+male | unknown speakers set | female-female | female+male | all   |
|-------|-----------|----------------------|---------------|-------------|-------|
|       |           |                      |               |             |       |
| *ADCNet | h=1       | 4.56                 | 4.13          | 8.85        | 8.12  |
|        | h=2       | 7.88                 | 8.32          | 9.53        | 9.01  |
|        | h=3       | 10.01                | 10.11         | 10.63       | 10.48 |
|        | h=4       | 7.94                 | 8.12          | 8.48        | 8.15  |
|        | h=5       | 2.21                 | 1.95          | 2.88        | 2.56  |
| ADCNet | h=1       | 4.78                 | 4.32          | 9.21        | 8.79  |
|        | h=2       | 8.26                 | 8.74          | 9.88        | 9.42  |
|        | h=3       | 10.17                | 10.34         | 11.03       | 10.75 |
|        | h=4       | 8.26                 | 8.59          | 8.82        | 8.34  |
|        | h=5       | 2.78                 | 2.51          | 4.01        | 3.45  |

TABLE 3. PESQ results with different \( h \) for DPCL with multi-head attention layer and normal K-means (*ADCNet) and DC with multi-head attention layer and canopy K-means (ADCNet) on unknown speakers in WSJ0-2mix.

| Model | male+male | unknown speakers set | female-female | female+male | all   |
|-------|-----------|----------------------|---------------|-------------|-------|
|       |           |                      |               |             |       |
| *ADCNet | h=1       | 1.26                 | 1.24          | 1.92        | 1.52  |
|        | h=2       | 1.73                 | 1.79          | 2.32        | 2.01  |
|        | h=3       | 2.52                 | 2.56          | 2.80        | 2.75  |
|        | h=4       | 1.49                 | 1.51          | 1.70        | 1.65  |
|        | h=5       | 1.13                 | 1.08          | 1.21        | 1.20  |
| ADCNet | h=1       | 1.28                 | 1.26          | 2.23        | 1.75  |
|        | h=2       | 1.77                 | 1.90          | 2.51        | 2.48  |
|        | h=3       | 2.58                 | 2.72          | 2.96        | 2.89  |
|        | h=4       | 1.77                 | 1.79          | 1.92        | 1.87  |
|        | h=5       | 1.20                 | 1.14          | 1.22        | 1.25  |

TABLE 4. SI-SNR (dB) results with different \( h \) for DPCL with multi-head attention layer and normal K-means (*ADCNet) and DPCL with multi-head attention layer and canopy K-means (ADCNet) on unknown speakers in WSJ0-2mix.

| Model | male+male | unknown speakers set | female-female | female+male | all   |
|-------|-----------|----------------------|---------------|-------------|-------|
|       |           |                      |               |             |       |
| *ADCNet | h=1       | 4.32                 | 4.04          | 8.61        | 8.15  |
|        | h=2       | 7.71                 | 8.22          | 9.34        | 9.08  |
|        | h=3       | 9.92                 | 9.87          | 10.55       | 10.53 |
|        | h=4       | 7.72                 | 8.05          | 8.34        | 8.23  |
|        | h=5       | 2.21                 | 1.83          | 2.72        | 2.51  |
| ADCNet | h=1       | 4.62                 | 4.38          | 9.12        | 8.52  |
|        | h=2       | 8.08                 | 8.73          | 9.74        | 9.46  |
|        | h=3       | 10.04                | 10.21         | 10.87       | 10.72 |
|        | h=4       | 8.17                 | 8.32          | 8.55        | 8.36  |
|        | h=5       | 3.68                 | 2.44          | 3.93        | 3.44  |

FIGURE 5. Example of audio spectrograms for separating a male utterance and a female utterance from the mixed speech. (a) is the spectrogram involving two speakers. (b) and (d) represents the signals of a female speaker and a male speaker respectively. (c) and (e) are the corresponding speeches separated by our model.

2) THE EFFECT OF THE NUMBER OF HEADS IN MULTI-HEAD SELF-ATTENTION LAYER

By observing the results of “*ADCNet” and “ADCNet” in Table 1-4, it can be discovered that with the increasement of \( h \), the performance of speech separation increases initially and decreases afterwards. When \( h = 3 \), the model performs the best. But when \( h = 5 \), SDR for *ADCNet is lower than the baseline DPCL model in Table 1, which indicates that multi-head self-attention can not perform well in all instances and the value of \( h \) is not the greater the better. Interpretations are given theoretically. Multi-head self-attention can acquire information from multiple aspects for feature extraction. It concatenates the output of self-attention and the input data are fully utilized. In addition, Highly optimized matrix multiplication can be achieved by dot-product in multi-head attention. As a result, compared with single attention layer,
the computational expense for multiple attention does not cost too much. So the model can perform well with the increasement of $h$ at first. However, when the value of $h$ is very large, much redundant information will be reserved and the size of parameters in model will also grows rapidly. Not only the model complexity is greatly increased, the computation resource is also used up. As a consequence, the model performance will become worse and worse. So appropriate value of $h$ can improve performance, while a big number will weaken the function of model.

3) COMPARISON WITH OTHER SEVERAL STUDIES

Our ADCNet is also compared with other models for single-channel speech separation on unknown speakers in WSJ0-2mix and WSJ0-3mix. As shown in Table 5, ADCNet achieves comparable performance with DPCL, uPIT, DANet and ADCNet models reported in [11], [13], [14] and [30]. Figure 5 shows the spectrograms for separating a male utterance and a female utterance from the mixed speech. The words for Sp1 and Sp2 are “We didn’t like that” and “A print media campaign will begin the following day” respectively. It can be observed that the spectrograms of original signals and their corresponding recovered speeches share many similarities, which indicates the effectiveness of our model.

V. CONCLUSION

In this paper, we propose combing the physiological characteristics of human beings with deep learning for single-channel speech separation. Specifically, we propose the joint optimization of multi-head self-attention and DPCL model. We prove that multi-head self-attention can achieve better results by experiments. What is more, we further optimize the model using density-based canopy K-means and our model outperforms other advanced models. For future work, we will investigate the label permutation problem and reduce the time delay for speaker-independent speech separation.

REFERENCES

[1] M. Cooke, J. R. Hershey, and S. J. Rennie, “Monaural speech separation and recognition challenge,” Comput. Speech Lang., vol. 24, no. 1, pp. 1–15, Jan. 2010.

[2] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” J. Acoust. Soc. Amer., vol. 25, no. 5, pp. 975–979, Sep. 1953.

[3] G. J. Brown and M. Cooke, “Computational auditory scene analysis,” Comput. Speech Lang., vol. 8, no. 4, pp. 297–336, 1994.

[4] E. B. D. Wang, G. J. Brown, and C. Darwin, “Computational auditory scene analysis: Principles, algorithms, and applications,” Acoust. Soc. Amer. J., vol. 124, no. 1, p. 13, 2008.

[5] M. N. Schmidt and R. K. Olsson, “Single-channel speech separation using sparse non-negative matrix factorization,” in Proc. 9th Int. Conf. Spoken Lang. Process., 2006, pp. 1–4.

[6] Y. M. Qian, C. Weng, X.-K. Chang, S. Wang, and D. Yu, “Past review, current progress, and challenges ahead on the cocktail party problem,” Frontiers Inf. Technol. Electron. Eng., vol. 19, no. 1, pp. 40–63, Jan. 2018.

[7] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, “Speech separation of a target speaker based on deep neural networks,” in Proc. 12th Int. Conf. Signal Process. (ICSP), Oct. 2014, pp. 473–477.

[8] S.-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.

[9] 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.

[10] M. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, “Joint optimization of masks and deep recurrent neural networks for monaural source separation,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Y. Luo and N. Mesgarani, “Com-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[16] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation,” ACM Trans. Graph., vol. 37, no. 4, pp. 1–11, Aug. 2018.

[17] N. Mesgarani and E. F. Chang, “Selective cortical representation of attended speaker in multi-talker speech perception,” Nature, vol. 485, no. 7397, pp. 233–236, May 2012.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998–6008.

[19] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.

[20] Z. Tan, M. Wang, J. Xie, Y. Chen, and X. Shi, “Deep semantic role labeling with self-attention,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 4929–4936.

[21] K. Ahmed, N. Shirish Keskar, and R. Socher, “Weighted transformer networks for machine translation,” 2017, arXiv:1711.02132. [Online]. Available: http://arxiv.org/abs/1711.02132.

[22] P. Verga, E. Strubell, and A. McCallum, “Simultaneously self-attending to words and phrases,” in Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol., vol. 1, 2018, pp. 872–884.

[23] T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, “Multi-head decoder for End-to-End speech recognition,” 2018, arXiv:1804.08050. [Online]. Available: http://arxiv.org/abs/1804.08050.
YANLIANG JIN received the B.S. and M.S. degrees in electrical engineering from Xidian University, Xi’an, China, in 1997 and 2000, respectively, and the Ph.D. degree in communication and information system from Shanghai Jiao Tong University, in 2005. He is currently an Associate Professorship with the School of Communication and Information Engineering (SCIE), Shanghai University (SHU). He has published more than 30 journal/conference papers. His research interests include mobile ad hoc networks (MANETs), wireless sensor networks (WSNs), wireless multimedia sensor networks (WMSNs), wireless broadband access, and signal processing.

CHENJUN TANG was born in Hubei, China, in 1995. She received the B.S. degree from Southern Medical University, Guangdong, China, in 2017. She is currently pursuing the M.S. degree in communication and information systems with the School of Communication and Information Engineering, Shanghai University. Her current research interests include speech separation and voiceprint recognition.

QIANHONG LIU was born in Henan, China, in 1993. She is currently pursuing the M.S. degree in communication and information systems with the School of Communication and Information Engineering, Shanghai University. Her current research interests include deep reinforcement learning and automatic driving.

YAN WANG was born in Anhui, China, in 1996. She is currently pursuing the M.S. degree in communication and information systems with the School of Communication and Information Engineering, Shanghai University. Her current research interests include indoor localization and ultra-wideband location.

**YANLIANG JIN**

Yanliang Jin et al.: Multi-Head Self-Attention-Based DPCL for Single-Channel Speech Separation

[24] M. India, P. Safari, and J. Hernando, “Self multi-head attention for speaker recognition,” 2019, arXiv:1906.09890. [Online]. Available: http://arxiv.org/abs/1906.09890

[25] C.-F. Liao, Y. Tsao, X. Lu, and H. Kawai, “Incorporating symbolic sequential modeling for speech enhancement,” in Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), Sep. 2019, pp. 2733–2737.

[26] J.-G. Sun, “Clustering algorithms research,” J. Softw., vol. 19, no. 1, pp. 48–61, Jun. 2008.

[27] M. Dianhui, “Improved Canopy-Kmeans algorithm based on mapreduce,” Jisuanji Gongcheng yu Yingyong (Comput. Eng. Appl.), vol. 48, no. 27, pp. 22–26, 2012.

[28] G. Zhang, C. Zhang, and H. Zhang, “Improved K-means algorithm based on density Canopy,” Knowl.-Based Syst., vol. 145, pp. 289–297, Apr. 2018.

[29] 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.

[30] Y. Luo, Z. Chen, and N. Mesgarani, “Speaker-independent speech separation with deep attractor network,” IEEE/ACM Trans. Audio, Speech, Language Process., vol. 26, no. 4, pp. 787–796, Apr. 2018.

[31] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., vol. 2, May 2001, pp. 749–752.

[32] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-channel multi-speaker separation using deep clustering,” in Proc. Conf. Int. Speech Commun. Assoc. (INTERSPEECH), 2016, pp. 545–549.

[33] 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.

[34] 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.