Listen, Think and Listen Again: Capturing Top-down Auditory Attention for Speaker-independent Speech Separation

Jing Shi[1,2,3*], **Jiaming Xu**[1,2*], **Guangcan Liu**[1,2,3], **Bo Xu**[1,2,3,4†]
[1]Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China
[2]Research Center for Brain-inspired Intelligence, CASIA
[3]University of Chinese Academy of Sciences
[4]Center for Excellence in Brain Science and Intelligence Technology, CAS. China
{shijing2014, jiaming.xu, liuguangcan2016, xubo}@ia.ac.cn,

Abstract

Recent deep learning methods have made significant progress in multi-talker mixed speech separation. However, most existing models adopt a driftless strategy to separate all the speech channels rather than selectively attend the target one. As a result, those frameworks may be failed to offer a satisfactory solution in complex auditory scene where the number of input sounds is usually uncertain and even dynamic. In this paper, we present a novel neural network based structure motivated by the top-down attention behavior of human when facing complicated acoustical scene. Different from previous works, our method constructs an inference-attention structure to predict interested candidates and extract each speech channel of them. Our work gets rid of the limitation that the number of channels must be given or the high computation complexity for label permutation problem. We evaluated our model on the WSJ0 mixed-speech tasks. In all the experiments, our model gets highly competitive to reach and even outperform the baselines.

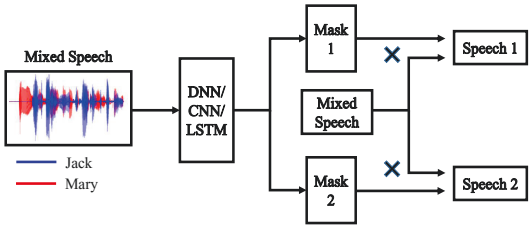
1 Introduction

Human auditory system gives us the extraordinary ability to converse in the midst of complex auditory scene. While completely intuitive and omnipresent in humans and animals alike, translating this remarkable ability into a quantitative model for computers remains a challenge [Elhilali, 2017]. Since its first description of this so-called cocktail party problem in 1953 by Colin Cherry [1953], many researchers have sought to the way to separate and recognize a mixed speech, especially in the case of single channel speech mixtures. Despite significant progress made in the recent years due to the success of deep learning, developing a computational auditory model to solve the cocktail party problem still has many unresolved issues, such as label ambiguity or permutation problem [Yu *et al.*, 2017] and output dimension mismatch problem [Chen *et al.*, 2017]. The former problem arises due to the fact that the order of the sources in the mixture is irrelevant, while the latter problem is usually encountered for an unfixed number of sources in the mixture.

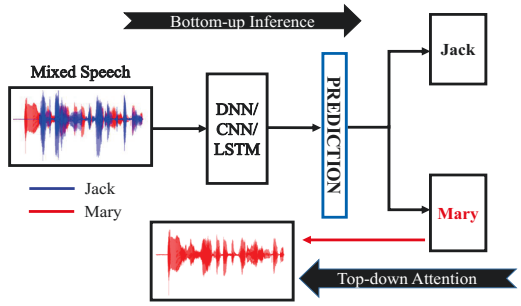
Recently, some researchers have attempted to alleviate these problems. To prevent the chaos brought by variant permutation, Yu et al. [2017] proposed a Permutation Invariant Training (PIT) method to pool over all possible permutations for N mixed sources ($N!$ permutations), and minimize the source reconstruction error no matter how labels are ordered. With the filtration process implemented inside, the network could be trained directly. However, PIT approach suffers the output dimension mismatch problem because it assumes a fixed number of sources and also suffers from its computation efficiency [Chen *et al.*, 2017]. In order to solve both permutation and output dimension problems, Hershey et al. [2016] proposed a Deep Clustering (DC) method which first maps the time-frequency units into an embedding space, and then generates a partition of the time-frequency units by employing a clustering algorithm, such as k-means. Following DC, Chen et al. [2017] proposed a Deep Attractor Network (DANet) which first forms k attractor points (cluster centers) in the embedding space and then pulls together the time-frequency units corresponding to the attractor points. Although DC and DANet are flexible to conduct speech separation on different number of sources in the mixture without retraining, both of them require a certain cluster number during evaluation to separate all the speech channels. Further, the algorithms adopted by DC and DANet are built on the Ideal Binary Mask (IBM) [Wang, 2005], which is so ideal and both of them use an additional binary weight to shield the silent region or emphasize the salient part.

To dig deep enough and trace to the source, the paradigm used in speech separation and the definition of related task get some limitations. From the perspective of human behavior in complex auditory scene, we never need to know how many sources are there in advance or to attend to all the channels together. Previous reports on dichotic listening behav-

*The first two authors contributed equally.
†Corresponding author



(a) Network for direct separation.



(b) Network with inference and stimuli-enhanced attention.

Figure 1: Structures of different networks for speech separation. Most existing methods build a direct regression structure to separate the mixed speech aimlessly, which causes a series of problems such as label ambiguity or output dimension mismatch. As a contrast, the network with inference and stimuli-enhanced attention gets the targets inferred from the mixtures to extract multiple channels.

ior [O'sullivan *et al.*, 2014; Cherry, 1953] show that human is not possible to listen to, and remember two concurrent speech streams, while listeners usually select the attended speech and ignore other sounds in the conditions where signals are either mixed or presented to separate ears. Simply speaking, rather than separating speech in complex auditory scene, human is more likely to extract targeted speech which interests him.

However, most of existing neural network based methods conduct the speech separation task with a straight pipeline. As illustrated in Figure 1(a), these baseline models actually propose a feedforward structure to conduct a multi-label regression process. In reality, rather than doing an aimless and all-channel-covered process, listeners get their priori knowledge to analyze mixed speech. Along with the physical properties of sounds in the environment, listeners exploit learned knowledge from their recent and lifelong experiences to further complement processing of complex auditory scenes [Bregman, 1990; Ciocca, 2008]. These learned "schemas" encompass a listener's familiarity with the statistical structure of sound sources, recent and long-term memories about specific sources, expectation about the state of the world (e.g., sounds produced through a human vocal tract), as well as their attentional state which helps steer brain processes towards targets of interest while ignoring background interferers.

Motivated by the ideas from behavior of auditory selective attention [Kaya and Elhilali, 2017] and "Biased Competition Theory" [Beck and Kastner, 2009] from cognitive science, we present a novel stimuli-enhanced speech separation framework. By "listening, thinking and listening again", our network achieves selectivity by jointing bottom-up inference and attention during the top-down loop. Figure 1(b) shows the basic process of this design. Our work gets several advantages: (1) Due to the variable number and definite aim from outputs of inference part, our method gets clear goal to extract aimed channel rather than ambiguous output. It overcomes the long-lasting permutation problem and output dimension mismatch problem in the baseline models. (2) Our inference-attention design is more fitted for the process about human behavior in complex auditory scene, which means a better interpretability. (3) Our network is fairly concise and easy to train. With limited computation complexity and basic model, our work still gets similar or better results than the baseline models.

2 Related Work

In recent years, due to the explosive development of removable devices, speech becomes the foremost entrance to access kinds of applications. Under this background, complex auditory scene analysis becomes a crucial problem to realize more robust intelligence. Within complex auditory scene analysis, speech separation is the bottommost task that algorithm takes the speech mixture as input to output different sources.

As mentioned in the former section, in our work, we raise a top-down attention framework to model the complex auditory scene. In fact, this framework could not only solve separation task towards human voice but also conduct many other related works. For example, if we adjust the top-down attention model from separating one's voice to noticing some specific phonemes, this framework could be viewed as a keyword-spotting system facing complicated auditory environment. Moreover, the top-down attention model could also be integrated with speech recognition techniques, which will give our framework a border functionality.

In this paper, we concentrate on separation task. In the following subsections, we will briefly introduce two directions that are closely related to the proposed research.

2.1 Speech Separation

As discussed before, in order to solve the cocktail party problem via computational auditory system, researchers have proposed many methods over the decades, from Computational Auditory Scene Analysis (CASA) [Brown and Cooke, 1994], Non-negative Matrix Factorization (NMF) [Schmidt and Olsson, 2006] to deep learning based approaches [Huang *et al.*, 2014; Yu *et al.*, 2017; Xu *et al.*, 2018]. As the most representative instance of dictionary learning, NMF decomposes each clean source into a set of speaker-dependent dictionaries and activations during training, and optimizes the activation for each source to achieve a global optimum. However, this method is restricted by many problems such as: (1) high computational complexity caused by the iteration algorithm, (2) unstable performance to deal with the noise-mixed or unknown speaker attended circumstance, (3) impractical attempting to reconstruct all possible sources even introducing group-sparsity penalty when the total number of available dictionaries is huge [Chen, 2017].

Several recent deep learning works [Yu *et al.*, 2017; Chen *et al.*, 2017] have also provided solutions. However, as discussed above, they also brought some problems like label ambiguity, permutation problem, computational complexity or

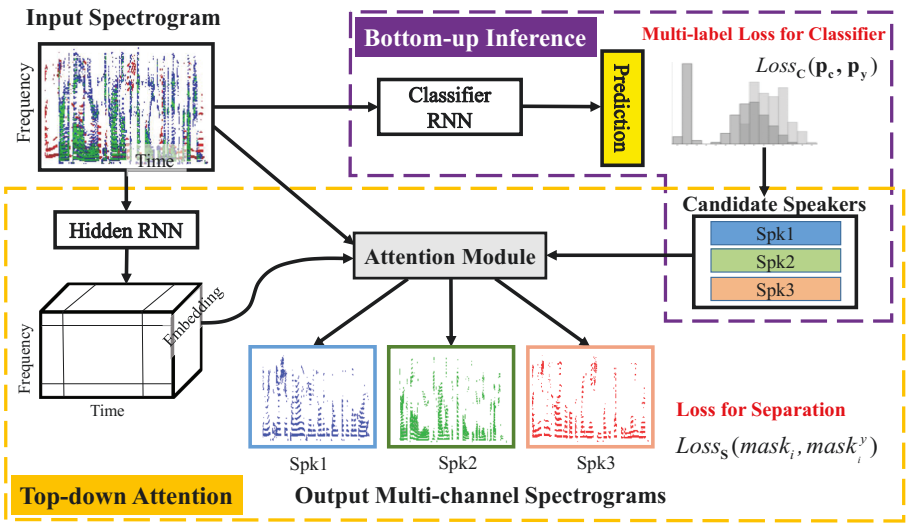


Figure 2: The framework of top-down auditory attention model for speech separation. This model constructs a loop with bottom-up inference and top-down attention process. From the former part, TDAA conducts the task to find the possible candidates to server as the targets for top-down process to extract multi-channel speeches towards each of them.

fixed known number of channels. And most of the problems are originated from the aimless separation process. To our delight, Xu et al. [2018] raised this problem and constructed a model with attention and memory to extract the speech of a talker. But their work could be conducted with only one given speaker, which is obviously impractical. Actually, the attentive target of one person in complex auditory scene is hard to follow, depending on not only auditory signal. In this work, we manage to model all the possible channels inferred from the mixture with only the speech information.

2.2 Top-down Attention

Recently, several attention-based neural network approaches have been built and demonstrated good performance in many areas with specialized neural architectures. Within these different networks, top-down attention also brought much promotion for many problems. Typically, visual attention is dominated by "goals" from our mind in a top-down manner, especially in the case of object detection. Cognitive science explains this in the "Biased Competition Theory" [Beck and Kastner, 2009], that human visual cortex is enhanced by top-down stimuli and non-relevant neurons will be suppressed in feedback loops when searching for objects [Cao *et al.*, 2015].

Similarly, auditory attention in cocktail party environment could also benefit from the top-down attention process. Human brain could construct recent and long-term memories about specific source of sound. With these memories, people could easily catch their interested source to adjust their attentional state which helps steer brain processes towards targets while ignoring background interferers. These processes are believed to play a crucial role in tackling the cocktail party problem because they impose constraints on the space of possible solutions. They can be viewed as top-down or feedback projections that control the system's performance to meet desired behaviors [Elhilali, 2017].

Based on above observations, integrating aim guided top-down attention into the computational auditory model would be a feasible solution for cocktail party problem. In our work, we build the bottom-up inference process to reason the possible interested sources as the aim and top-down attention process to extract speech channel for each single one.

3 Top-down Auditory Attention Model

In our work, we use Top-Down Auditory Attention model (hereinafter dubbed TDAA) to conduct the selectivity of speakers and separation of their speech when facing complex auditory environment. The framework of TDAA model consists of two main parts: bottom-up inference and top-down attention. The following chapter will describe the whole model and the two critical modules carefully.

3.1 General Framework

To put it sequentially, TDAA model works following a pipeline from "Listening" to "Thinking" and "Listening again". We mimic the human behavior that human may focus to one specific object within a complicated auditory scene after a first short time to decide which aim to listen attentively. At the very first time, the bottom-up inference happens. In our model, we utilize this procedure to output an accurate candidate speakers. After that, each candidate servers as the query to promote the following top-down attention process.

As described in Figure 2, given a spectrogram $X_{t,f}$ (transformed from original mixture speech x by Short-Time Fourier Transformation), we first use a Bidirectional Long-Short Term Memory (BiLSTM) to classify this input mixture. As the mixture speech gets several speakers, the classifier will make prediction as a multi-label classification problem. Assume that the number of all known speakers (speakers in training dataset) is N, then we make the aim distribution $\mathrm{P}_y \in \mathbb{R}^N$ of one speech as a vector consists of 1 or 0. The "1" element means the corresponding speaker is indeed in this mixture speech. The bottom-up inference process actually is to build a neural network to make prediction $\mathrm{P}_c \in \mathbb{R}^N$ about

the likely speakers from the mixture speech. Here the loss of this classifier $Loss_C(\mathrm{P}_C, \mathrm{P}_y)$ will be used to optimize this multi-label classification problem. As the prediction generates kinds of different results, the number of predicted candidate speakers varies according to different mixtures.

After getting the candidate speakers, the top-down attention module will take the candidates as queries to extract related speech channel for each of them. Actually, as mentioned above, human is not possible to listen to, and remember two concurrent speech streams. Moreover, the interesting channel to attract one's attention depends on a complex set of factors, including but not limited to visual, auditory and even mental activities. Different from [Xu *et al.*, 2018] that the aimed speaker is single and artificially given, we manage to infer and cover all the possible channels for one to focus from purely auditory inputs.

3.2 Attention Module

With the input mixture spectrogram $X_{t,f}$ (t gets a total length of T and f gets F frequencies altogether) and aimed candidates, a concise and reliable attention network is executed. For the mixture $X_{t,f}$, TDAA first adopts a Bidirectional Long-Short Term Memory to map the input into hidden states $H_{t,m}$, where the m equals number of hidden units in this layer. Following that, a feed-forward layer embeds the hidden states to form an embedding matrix whose every time-frequency unit $h_{t,f} \in \mathbb{R}^d$ implicitly represents features in hidden states of the unit t, f from mixture spectrogram.

For each candidate i, a hidden embedding layer servers as a look-up matrix $E \in \mathbb{R}^{N\times d}$ to choose his speaker embedding E_i from it. Here we use the same dimension number as $h_{t,f}$. Then, attention module takes $h \in \mathbb{R}^{T\times F\times d}$ and E_i to compute $mask_i \in \mathbb{R}^{T\times F}$ as simulation to Ideal Ratio Mask (IRM) $mask_i^y \in \mathbb{R}^{T\times F}$ for channel i in mixture. Towards each unit $h_{t,f} \in \mathbb{R}^d$ and speaker E_i, kinds of different attention mechanisms could be used, such as $h_{t,f}E_i$ (dot production), $h_{t,f}WE_i$ (general), and $gtanh(W[h_{t,f}; E_i])$ (cont) [Luong *et al.*, 2015] , where W and g are all learned parameters. In our work, we take the simplest dot production as attention mechanism in all the experiments.

3.3 Recurrent Inference Process

We all know that everyone gets a unique voiceprint, but different speakers may be very similar in their voices. Perhaps due to this characteristic, although trained from a fixed set of known speakers, TDAA is still able to extract unknown speakers. This good performance of generalization has also been observed from [Li *et al.*, 2017; Arik *et al.*, 2017]. To describe it from the angle of algorithm, our model takes the voiceprints of known speakers as "seeds" to approximate diverse talkers no matter they are in training dataset or not.

The experiment results show that bottom-up inference module works well to predict known speakers. However, for the unknown condition, we found the Classifier RNN sometimes outputs candidates referring to same speaker as follows:

Prediction:	Spk2,Spk5→Unk1	Spk8→Unk2

We call this phenomenon as "Repeated Prediction Problem", and more information about this problem could be seen from Appendix A. This disadvantage may originate from insufficient training speakers. Limited samples results in several seeds to respond to only one speaker. To overcome this problem, we raise a recurrent attention inference method to get rid of repeated prediction.

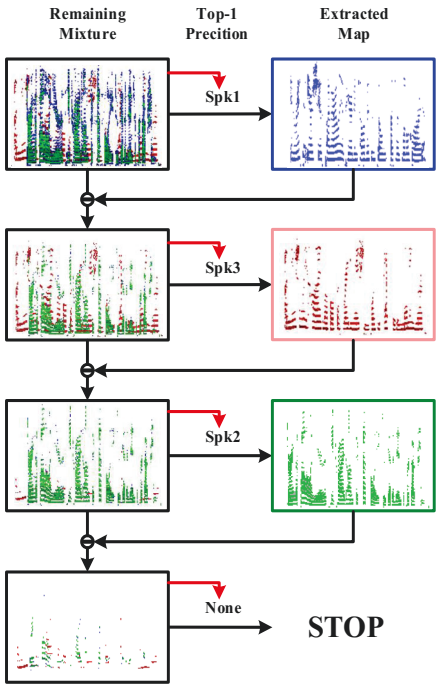


Figure 3: The course of recurrent inference process with multiple steps. After getting the trained framework, we use this recurrent structure to predict the most salient speaker at each step.

As illustrated in Figure 3, the bottom-up inference module could be conducted as a recurrent process. In each step, we use the Classifier RNN to output the top-1 prediction result rather than several candidates. After getting the most likely attention aim, single-channel top-down attention will extract his speech from current mixture. Following that, the mixture will be replaced by remaining spectrogram. This recurrent process stops once current mixture predicts no talker left.

Recurrent attention inference enforces the module to concentrate on the most salient speaker so as to avoid Repeated Prediction Problem. Just like human brain's ability to focus towards targets of interest while ignoring background interferers, a diminishing mixture mechanism could also weakened the interruption from previous candidate. Through this process, TDAA could figure out how many channels are there precisely even in mixture with all unknown speakers.

3.4 Loss Function Definition

The loss function for speech separation task has been attempted by many works [Wang *et al.*, 2016]. Similar to the intention behind those models, TDAA also makes some adjustment to boost performance. In our work , the loss function is defined as follows:

$$Loss = Loss_{\mathrm{C}}(\mathrm{p_c}, \mathrm{p_y}) + Loss_{\mathrm{S}}(mask_i, mask_i^y) \qquad (1)$$

Setting	Acc	top-2 Recall	top-3 Recall
2 speakers	96.8	84.1	86.9
2 speakers*	97.2	84.9	87.2
1,2,3 speaker(s)	97.5	88.6	92.7
1,2,3 speaker(s)*	**97.8**	**90.4**	**94.4**

Table 1: Accuracy rate (%) and recall rate (%) with different setup to train the classifer RNN from bottom-up inference process. The " * " here means the existence of Summarization Normalization loss.

$$Loss_{\mathrm{C}} = MSE(\mathrm{p_c},\mathrm{p_y}) + \alpha \cdot MSE\left(1, \sum_{j}^{N} \mathrm{p_c^j}\right) \tag{2}$$

$$Loss_{\mathrm{S}} = \sum_{i \in Cadidates} MSE(mask_i, mask_i^y) + \beta \cdot MSE(1 \in \mathbb{R}^{T \times F}, \sum_i mask_i) \tag{3}$$

The $Loss_C$ above is the loss for bottom-up classifier while the $Loss_S$ means the loss for separation. Besides the standard definition for the classifier or minimum error between the predicted masks and true masks, we add another two similar items to regularize the training. We call the later part in $Loss_{\mathrm{C}}$ and $Loss_{\mathrm{S}}$ as "Summarization Normalization". This design encourages the networks to output with more diversity and exclusiveness in their elements.

4 Experiments

4.1 Dataset and Setup

Actually, our model builds a unified framework to process kinds of different tasks towards complex auditory scene. Through the two phases of inference and attention, many other typical tasks, such as spoken term detection and speech recognition, could also benefit from this paradigm. In this paper, we only concentrate on single-channel speaker-independent speech separation task.

We evaluated TDAA on speech mixture with different number of speakers. Data from Wall Street Journal (WSJ0) corpus has been used in our work. For condition of two-speaker, the WSJ0-2mix dataset was introduced in [Hershey *et al.*, 2016]. The 30h training set and the 10h validation set contains two-speaker mixtures generated by randomly selecting speakers and utterances from the WSJ0 training set *si_tr_s*, and mixing them at various Signal-to-Noise Ratios (SNRs) uniformly chosen between 0 dB and 5 dB. The 5h test set was similarly generated using utterances from 18 speakers from the WSJ0 validation set *si_dt_05* and evaluation set *si_et_05*. For three-speaker experiments, similar methods were adopted while the number of speakers was three. In the validation set, TDAA could be used to evaluate the source separation performance on known talkers, which is the so-called Closed Conditions (CC) in [Hershey *et al.*, 2016; Isik *et al.*, 2016]. As a contrast, Open Condition (OC) will provides unknown speakers to conduct from test set. In all our experiments, we use the whole mixture as inputs to train our model and test on it. Further details on the training setup are given in Appendix B.

To quantitatively evaluate results and compare with other published works, we report the overall performance via the Signal-to-Distortion Ratio (SDR) [Vincent *et al.*, 2006] metric. SDR measures speech enhancement performance in speech related tasks which use the original single speaker's speech as target. However, under the background of Ideal Ratio Mask speech separation methods, many works just predict magnitude for each channel and directly use phase of the mixed speech to recover original source. Definitely, this approximation takes systemic error. But for human, these recovered speeches make very little difference in speech intelligibility, especially when the number of channels is few. Based on this consideration, we also use the speech recovered from the IRM magnitude and mixed phase to server as aimed speech to calculate the $\widetilde{SDR}$, which is a good supplement to analyze the results and verify the effectiveness of training.

Model	2mix CC SDR\$\widetilde{SDR}$	2mix OC SDR\$\widetilde{SDR}$
Oracle NMF [Hershey *et al.*, 2016]	5.1	-
CASA [Hershey *et al.*, 2016]	2.9	3.1
DPCL [Hershey *et al.*, 2016]	5.9	5.8
DPCL+ [Hershey *et al.*, 2016]	-	10.3
PIT-DNN [Yu *et al.*, 2017]	5.7	5.6
PIT-CNN [Yu *et al.*, 2017]	7.7	7.8
TDAA top-1	**9.1** \ **12.6**	7.5 \ 9.9
TDAA top-2	8.5 \ 12.1	4.1 \ 7.0
TDAA top-5*	9.0 \ 12.4	8.1 \ 11.6
TDAA top-5to2	8.6 \ 11.9	**7.8** \ **11.2**
TDAA GT*	9.4 \ 12.8	-
IRM	12.3	12.5

Table 2: SDR improvements (dB) for different separation methods on the WSJ0-2mix dataset. As mentioned in Section 4.1, $\widetilde{SDR}$ means the SDR improvements with IRM as target. The model with " * " provides results from ideal condition (which is not true), and the numbers in bold show the true best results with TDAA.

4.2 Bottom-up Inference Performance

This part shows the performance of reasoning in bottom-up inference process. At this stage, TDAA mainly uses Classifier RNN to predict possible candidates. Table 1 shows the results of basic model. In this process, we test different settings to train the network to predict, such as number of speakers in training mixtures and classification loss Summarization Normalization. The training and test speakers to mix are all from WSJ0 training set, but the samples are not overlapped. And the test set is limited to mixtures with 2 speakers. Due to the multi-label classification problem in actually, here we describe 3 metrics to evaluate the performance. Accuracy here means the proportion of correct prediction to all known speakers (101 speakers in WSJ0 training set). The top-2 recall rate is calculated using the first 2 predicted speakers from our model as output while the top-3 using the first 3 ones.

From the table we can make several observations. First, the accuracy towards all the elements are considerably high with our base Classifier RNN. Actually, every mixture speech gets a fairly unbalanced ground-truth label with just 2 positive element. A simple model to predict all negative results is easily to get an accuracy about 98%. In reality, this unbalance data indeed brings problems for TDAA to train, and we adopted

Bottom-up	top-2 Recall	top-3 Recall
	93.2(**+2.8**)	97.7(**+3.3**)
Top-down	2mix CC SDR	2mix OC SDR
	9.2(**+0.7**)	7.7(**+3.6**)

Table 3: Results of recurrent inference attention in bottom-up and top-down process respectively. Figures in brackets mean the improvement over basic model without recurrent inference.

a relatively small learning rate to alleviate it. Second, not surprisingly, the training data with variable number of speakers brings promotion of the results. This conclusion is also in line with the condition from [Isik *et al.*, 2016]. Third, the prediction network could benefit from the Summarization Normalization that it weakens problems from the similar speaker from training set. Finally, the top-3 prediction gets obvious promotion over top-2 recall rate, which also reveals the problem of repeated prediction.

From these experiments, we could conclude that our bottom-up inference process is an effective method to catch the aimed candidates towards mixed speech. Through this stage, our following attention mechanism gets a satisfying set of aimed candidate speakers to extract their channels.

4.3 Speech Separation Performance

After getting the candidates, our top-down attention process could be used to conduct the speech separation task. In this section, we take the WSJ0-2mix datasets to evaluate our model. To compare equally, here we set the number of channels to output to be known as 2, but actually TDAA could work with no need for this condition. Corresponding to Table 2, there are several different settings to output the final separation speech after getting trained framework. TDAA top-1 model means to extract the mask for the most likely candidate and the $1 - mask$ for the other channel. For the standard TDAA top-2 model, we use the top-2 speaker predicted from bottom-up inference process as the aim to extract their speeches respectively. For the looser condition, we use the top-5 candidates to produce 10 most likely speaker pairs to choose the one with best SDR (which is not true). Following that, we use the top-5 candidates to find two of them with least cosine similarity without supervision (TDAA top-5to2). In addition, for the known speakers in CC, the two ground-truth speakers servers as the candidates to attend (TDAA GT).

From above different setting, we could get some cognition about the upper bound and limitation towards TDAA. For the top-2 model and ground-truth model, the latter one shows upper limit with the trained top-down attention and the gap between them indicates the interference brought by the wrong prediction from bottom-up inference process. Similarly, the gap between top-2 and top-5 model shows the deficiency from the trained multi-label classifier. Based on this observation, we adopt another simple method to extra two candidates from top-5 predicted ones as the final output. As a reference, we also provided the IRM result which is the oracle and upper bound achievable on this task.

The results show that our top-down attention process works well on extracting each speech channel following the bottom-up stage. For the known speakers, basic top-2 model could reach a satisfying performance while open condition should be settled with other settings. In addition, with kinds of different settings, the $\widetilde{SDR}$ metric shows almost the consistent trends with the standard SDR results, which indicates the effectiveness for out model to match the IRM magnitude and a good result of intelligibility.

Method	MS-CC	MS-OC
Oracle NMF	4.4	-
DC local	3.5	2.8
DC global	2.7	2.2
TDAA recurrent attention	**4.6**	**3.0**

Table 4: SDR improvement (dB) for mixtures of three speakers.

Compared with other art methods, under the same condition that the number of channels is given, TDAA can achieve similar and better performance than the original DPCL [Hershey *et al.*, 2016] and DNN or CNN based PIT [Yu *et al.*, 2017] especially in closed condition, but underperforms the more complicated DPCL+ [Isik *et al.*, 2016]. We should point that TDAA adopts basic networks such as original bidirectional LSTM and simplest dot production attention algorithm, and we did not fine-tune our settings or add some more complicated mechanism such as recurrent dropout, norm constraint used in DPCL+. In addition, the results of DC based methods are quite heavily influenced by the background noise threshold which is a tricky technology.

4.4 Recurrent Inference Model Performance

Basic attention model has shown an obvious advantage over known speakers. The deficiency in open condition mainly originates from the bottom-up candidates. After observing the generated output, we find that the two predicted speeches sound pretty alike in many mixtures. This is also to say that towards many unknown speakers' mixtures, the bottom-up inference process tends to predict two speaker towards the same single one. As mentioned above, we call this phenomenon as "Repeated Prediction Problem".

Although we could enlarge the candidates scope and execute some filtration to get the final output like the TDAA top-5to2 model, but this method is a bit tricky and restricted. To attenuate this problem, we manage to improve the bottom-up process with recurrent inference attention model to extract candidates step by step. Figure 3 describes this course. If we set the process to stop after two steps, we could get the results to compare the experiments in above sections.

As illustrated in Table 3, although with an identical network and parameters, the performance in bottom-up and top-down process both gets a promotion after using the recurrent inference method. More importantly, recurrent inference attention could stop at the second step with a success rate over 87%. When the number of mixed speakers is unknown, this method could proceed the separation pipeline spontaneously and stop at an appropriate moment. Table 4 also shows the results on 3 speakers mixtures as the setting from MS-CC and MS-OC introduced in [Isik *et al.*, 2016]. The result proves a good generalization and adaptation ability of recurrent inference attention in TDAA towards different setup.

5 Conclusion and Future Works

In this paper, we present a novel framework with bottom-up inference and top-down attention modules to conduct mixed speech related problems when facing complex auditory scene. Different from direct separation methods in prior arts, our method builds a new loop to extract sources from predicted target speakers rather than aimless regression.

Several advantages could be brought by TDAA model. First, owing to the design of bottom-up inference, the number of channels in mixture could be variable and even unknown, which we believe is a big step to process speech data from complicated real environment. Second, with just basic network, simplest attention function, and spectral magnitude only, TDAA reaches or outperforms baseline models. Further tuning and more complicated networks may lead to more promotion in performance. Third, because TDAA gets a highly abstract structures, various tasks involving complex auditory scene could be easily integrated in our framework by defining different top-down module.

Although the performance by our model reaches a good level, there still are some challenging problems. For instance, the method to take the known speakers as the "seeds" to approximate unknown talker exists some systematic errors, which may cause the gap between the open-condition and closed-condition. In future works, we intend to conduct further research on the human behavior to alleviate or fix the problem brought by the unknown speakers. In addition, the mechanism to attract one's attention in complex scene could also be redesigned with kinds of different originations, such as visual signal or spontaneous attention, which could bring more usefulness for our model to fit into real and multi-modal environment. Moreover, existing methods to conduct speech separation are built based on the artificially mixed data, which may simplify the problem. In future, we would explore the method to handle true condition with less supervision.

Acknowledgements

We thank the reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (61602479, 91720000), the Independent Deployment Project of CAS Center for Excellence in Brain Science and Intelligent Technology (CEBSIT2017-02) and Advance Research Program (6140452010101).

A Repeated Prediction Problem

In our work, the Classifier RNN in bottom-up inference module is in charge of predicting the possible candidates for the following top-down attention to work with these "seeds". Given the known speakers, TDAA works pretty well to find the aim speakers. However, when the speakers in mixtures have never been seen in training process, the repeated prediction problem may arise.

In reality, although everyone has a unique voiceprint, there are also many very similar vocals. To be specific, the blue circles in Figure 4 shows the distribution in 2-D PCA space for all the embeddings of known speakers in WSJ0 dataset trained from our top-down attention model. From this figure

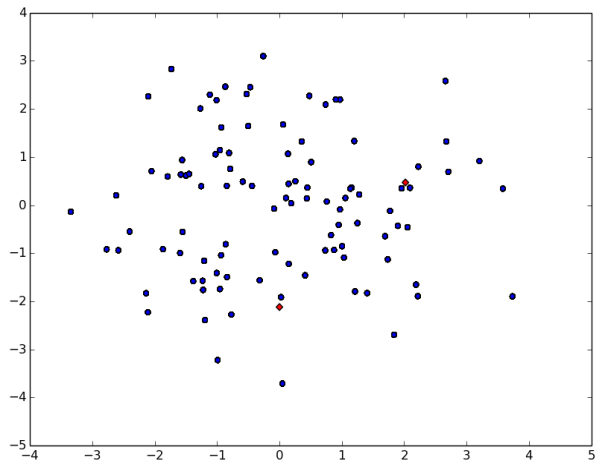


Figure 4: The 2-D Principal Component Analysis (PCA) distribution of trained embeddings for all the known speakers (blue circle) in WSJ0 datasets and two unknown speakers (red diamond).

we could observe that there are even speakers whose embeddings are so close that it is hard to discriminate them. Besides, there are two unknown speakers with red diamond in this figure and one of them (the upper one) has a similar voiceprint with two known speakers. Under this circumstance, the Classifier RNN in bottom-up inference prefers to predict the top-2 candidates toward this single unknown speaker. Because of this reason, the basic model with restricted top-2 candidates gets a poor performance over Open Condition (OC).

The recurrent inference method used in our work is designed to fix this problem. With this method, the bottom-up inference process could be finished in a recurrent strategy. At each step, the candidates predicted from former steps are excluded in order to weaken the influence brought by the similar known speakers. From the experiment results, we could conclude this recurrent inference method is effective to conduct repeated prediction problem.

B Training Setup in Details

In our experiments, all data are resampled to 8 kHz and the max length of mixture is restricted to 5 seconds to reduce computational and memory costs. The magnitude spectra is served as input feature, computed using Short-Time Fourier Transform (STFT) with 32 ms window length, 16 ms hop size and the sine window. We randomly generate 16 samples for one mini-batch. And we use a random shift of the frames to augment the training data. Our models were trained using Nesterov Adam with a decreasing learning rate of λ from 10^{-4} to 10^{-6}.

For the architecture of networks, we use a 3-layer BiLSTM with 300 hidden units and a following feed-forward layer to serve as Classifier RNN. Actually, different number of layers with 2 or 4 makes little difference in our experiments, and we choose the balanced one. As for Hidden RNN, a 4-layer BiLSTM with 300 hidden units was adopted. The dimension of Embedding $h_{t,f} \in \mathbb{R}^d$ and speaker vector E_i is set to be identical, which is actually not necessary, and equal to 50 in our model. Dropout operation was used in each feed-forward

layer with rate of 50%.

For the loss function, in Equation 2, due to the different range of loss for the classifier (the first item) and the Summarization Normalization (the latter item) at initialization, we set α to be zero in first 20 epoches and 0.002 after that to harmonize these two items. Similarly, β in Equation 3 is set to zero from beginning and 0.1 after 20 epoches. Other loss functions could also be attempted, such as cross-entropy loss, Kullback-Leibler divergence and so on. We have tested some different choices, but MSE showed the best result.

References

[Arik *et al.*, 2017] Sercan Arik, Gregory Diamos, Andrew Gibiansky, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. Deep voice 2: Multi-speaker neural text-to-speech. *arXiv preprint arXiv:1705.08947*, 2017.

[Beck and Kastner, 2009] Diane M Beck and Sabine Kastner. Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision research*, 49(10):1154–1165, 2009.

[Bregman, 1990] AS Bregman. Auditory scene analysis: The perceptual organization of sound. 1990, 1990.

[Brown and Cooke, 1994] Guy J Brown and Martin Cooke. Computational auditory scene analysis. *Computer Speech & Language*, 8(4):297–336, 1994.

[Cao *et al.*, 2015] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

[Chen *et al.*, 2017] Zhuo Chen, Yi Luo, and Nima Mesgarani. Deep attractor network for single-microphone speaker separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 246–250. IEEE, 2017.

[Chen, 2017] Zhuo Chen. *Single Channel auditory source separation with neural network*. PhD thesis, Columbia University, 2017.

[Cherry, 1953] E Colin Cherry. Some experiments on the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America*, 25(5):975–979, 1953.

[Ciocca, 2008] Valter Ciocca. The auditory organization of complex sounds. *Frontiers in bioscience: a journal and virtual library*, 13:148–169, 2008.

[Elhilali, 2017] Mounya Elhilali. *Modeling the Cocktail Party Problem*, pages 111–135. Springer International Publishing, Cham, 2017.

[Hershey *et al.*, 2016] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[Huang *et al.*, 2014] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Deep learning for monaural speech separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 1562–1566. IEEE, 2014.

[Isik *et al.*, 2016] Yusuf Isik, Jonathan Le Roux, Zhuo Chen, Shinji Watanabe, and John R Hershey. Single-channel multi-speaker separation using deep clustering. *arXiv preprint arXiv:1607.02173*, 2016.

[Kaya and Elhilali, 2017] Emine Merve Kaya and Mounya Elhilali. Modelling auditory attention. *Phil. Trans. R. Soc. B*, 372(1714):20160101, 2017.

[Li *et al.*, 2017] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu. Deep speaker: an end-to-end neural speaker embedding system. *arXiv preprint arXiv:1705.02304*, 2017.

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *empirical methods in natural language processing*, pages 1412–1421, 2015.

[O'sullivan *et al.*, 2014] James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral Cortex*, 25(7):1697–1706, 2014.

[Schmidt and Olsson, 2006] Mikkel N Schmidt and Rasmus Kongsgaard Olsson. Single-channel speech separation using sparse non-negative matrix factorization. In *Spoken Language Proceesing, ISCA International Conference on (INTERSPEECH)*, 2006.

[Vincent *et al.*, 2006] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006.

[Wang *et al.*, 2016] Guan Xiang Wang, Chung Chien Hsu, and Jen Tzung Chien. Discriminative deep recurrent neural networks for monaural speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2544–2548, 2016.

[Wang, 2005] DeLiang Wang. On ideal binary mask as the computational goal of auditory scene analysis. *Speech separation by humans and machines*, pages 181–197, 2005.

[Xu *et al.*, 2018] Jiaming Xu, Jing Shi, Guangcan Liu, Xiuyi Chen, and Bo Xu. Modeling attention and memory for auditory selection in a cocktail party environment. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[Yu *et al.*, 2017] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 241–245. IEEE, 2017.