The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Modeling Attention and Memory for Auditory Selection in a Cocktail Party Environment

Jiaming Xu,[1,2] **Jing Shi,**[1,2,3] **Guangcan Liu,**[1,2,3] **Xiuyi Chen,**[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
{jiaming.xu, shijing2014, liuguangcan2016, chenxiuyi2017, xubo}@ia.ac.cn

Abstract

Developing a computational auditory model to solve the cocktail party problem has long bedeviled scientists, especially for a single microphone recording. Although recent deep learning based frameworks have made significant progress in multi-talker mixed speech separation, most existing deep learning based methods, focusing on separating all the speech channels rather than selectively attending the target speech and ignoring other sounds, may fail to offer a satisfactory solution in a complex auditory scene where the number of input sounds is usually uncertain and even dynamic. In this work, we employ ideas from auditory selective attention of behavioral and cognitive neurosciences and from recent advances of memory-augmented neural networks. Specifically, a unified Auditory Selection framework with Attention and Memory (dubbed ASAM) is proposed. Our ASAM first accumulates the prior knowledge (that is the acoustic feature to one specific speaker) into a life-long memory during the training phase, meanwhile a speech perceptor is trained to extract the temporal acoustic feature and update the memory online when a salient speech is given. Then, the learned memory is utilized to interact with the mixture input to attend and filter the target frequency out from the mixture stream. Finally, the network is trained to minimize the reconstruction error of the attended speech. We evaluate the proposed approach on WSJ0 and THCHS-30 datasets and the experimental results demonstrate that our approach successfully conducts two auditory selection tasks: the top-down task-specific attention (e.g. to follow a conversation with friend) and the bottom-up stimulus-driven attention (e.g. be attracted by a salient speech). Compared with deep clustering based methods, our method conducts competitive advantages especially in a real noise environment (e.g. street junction). Our code is available at https://github.com/jacoxu/ASAM.

Introduction

Cocktail party problem describes human's ability that listeners can easily attend to one speaker in a multi-speaker environment (O'sullivan et al. 2015). Since its first description in 1953 by Colin Cherry (1953), many researchers have sought to understand and model the selective attention process of multi-talker speech. Despite the 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, Luo, and Mesgarani 2017). The former problem raises due to the fact that the order of the sources in the mixture is irrelevant, while the latter problem is usually encountered by a unfixed number of sources in the mixture.

Recently, some researchers have attempted to alleviate these problems. For example, Yu et al. (2017) proposed a Permutation Invariant Training (PIT) method to solve the permutation problem by pooling over all possible permutation for N mixed sources ($N!$ permutations), and minimize the source reconstruction error no matter how labels are ordered. 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 a 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 and produce the state-of-the-art separation, both of them require a certain cluster number during evaluation to separate all the speech channels.

Obviously, this assumption to be given a certain cluster number is too strict to offer a satisfactory solution in a complex auditory scene, where it is difficult to determine the number of the input sources. From the previous reports on dichotic listening behavior (O'sullivan et al. 2015; Cherry 1953), human is not able 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. Such manner makes human auditory system have the ability to attend their interesting speech in a complex auditory scene without considering the number of the input auditory signals. Thus, developing an auditory attentive selection model to solve the cocktail party problem may

*Corresponding author

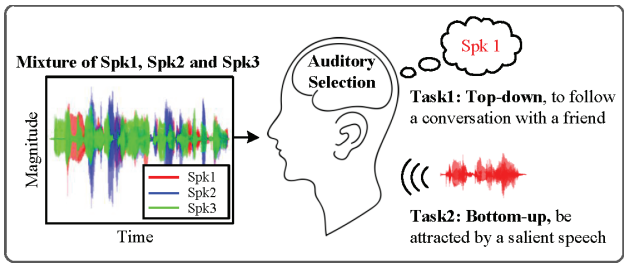

Figure 1: Two specific attention tasks for auditory selection in a three speech mixture environment. One is top-down task-specific attention, and the other is bottom-up stimulus-driven attention.

be more practical.

Besides, the speaker identity is not fully utilized (or unintentionally ignored) in most existing methods which only use the speaker identity to guide the label permutation. Actually, listener in a cocktail party environment always first tries to determine the identity of the target speaker with the help of visual perception or other information. If the listener is familiar with the acoustic feature of the target speaker, the corresponding acoustic memory would be extracted from brain circuits directly to assist auditory attention based on the speaker identity. While for a unknown target speaker, the listener should first perceive the acoustic feature from a salient speech and then update the memory corresponding with the new speaker identity for the following continuous auditory attention.

Following the ideas from auditory selective attention of behavioral and cognitive neurosciences (Kaya and Elhilali 2017), two specific auditory attention tasks are illustrated in Figure 1. Assume there are three speakers, Spk1, Spk2 and Spk3, in a cocktail party environment. Task 1 shows a top-down task-specific attention which requires the listener to follow a conversation with a friend based on a given speaker identity, such as Spk1. While task 2 is a bottom-up stimulus-driven attention where the listener's attention is attracted by a given salient speech of the target speaker. Both of them may usually work simultaneously in human's auditory pathway, or the bottom-up attention would be transformed into a top-down attention after a few interactions. To make the comparison clearer, we decompose the auditory attention into two separate tasks: top-down and bottom-up, but integrate them into one unified framework.

Specifically, this paper propose a unified Auditory Selection framework by modeling Attention and Memory (abbr. to ASAM). Our method first accumulates the prior knowledge (that is the acoustic feature to one specific speaker) into a life-long memory which is one particular external memory module without the need to reset it during training, meanwhile a speech perceptor is trained to extract the temporal acoustic feature and update the memory online when a salient speech is given. Then the learned memory is utilized to interact with mixture input to attend and filter the target speech. Our main contributions are three-fold:

(1) To our best knowledge, this is the first time to integrate top-down task-specific attention and bottom-up stimulus-driven attention into one unified computational auditory framework, which is closer to human auditory behavior, fully trainable and easy to implement without any specific settings, such as giving a certain number of the input auditory signals during evaluation or setting a threshold value to ignore the background noises.

(2) We exploit a life-long memory following one speech perceptor to accumulate the prior knowledge during training, which make our model have the ability to recall the acoustic memory of the trained speaker or perceive the salient speech stimulation in our unified framework, and even update the memory online once the speaker identity is determined after salient speech stimulation.

(3) We test and verify experiments based on two publicly available speech datasets: WSJ0 and THCHS-30. The various experimental results demonstrate that our approach successfully conduct two auditory selective attention tasks, and even show the robustness in a real noise environment.

The remainder of this paper is organized as follows: In Section 2, we briefly survey several related works. Section 3 describes the proposed framework ASAM and gives implementation details. Experiments are presented in Section 4. Finally, conclusions are given in the last Section.

Related Work

In the following subsections, we briefly introduce two directions that are related to the proposed research.

Speech Separation and Auditory Selection

As discussed before, in order to solve the cocktail party problem via computational auditory system, researchers have proposed many speech separation 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). NMF, as the most representative instance of dictionary learning, 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, these decomposition based methods mainly have the following limitations: (1) Most of these methods achieve a global optimum through iterative method, which usually results in high computational complexity during evaluation. (2) The noise or unknown speaker in the background would prevent the decomposition model to achieve high quality separation, even though the target dictionary has been pre-learned well. (3) When the total number of available dictionaries is huge, attempting to reconstruct all possible sources is impractical even introducing group-sparsity penalty (Chen 2017). A prior knowledge to choose a meaningful subset of the pre-learned dictionaries is helpful for decomposition, however, in a complex auditory scene, the input sources are usually uncertain and even dynamic. Although the recent Deep Clustering (DC) (Hershey

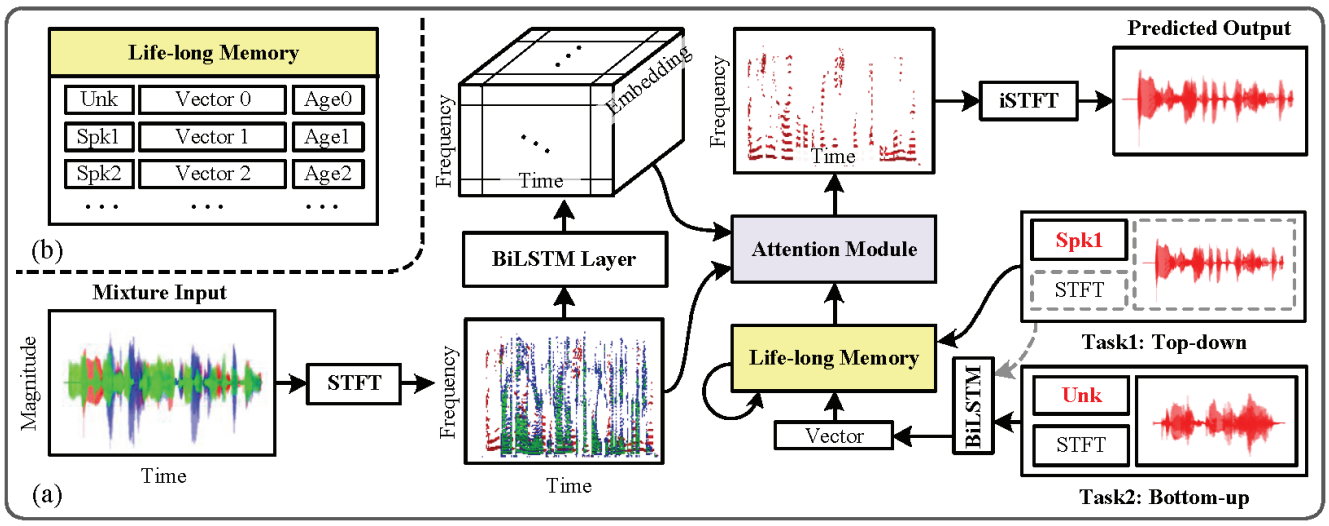

Figure 2: An illustration of our Auditory Selection with Attention and Memory (ASAM). (a): The overall architecture of the proposed ASAM. (b): Life-long memory module to memory the prior knowledge. In top-down attention scene, the dashed boxes and arrow are only conducted in the training phase and removed in the evaluation time.

et al. 2016) and Deep Attractor Network (DANet) (Chen, Luo, and Mesgarani 2017) have overcome the label permutation problem and output dimension mismatch problem. Both of them also do not move away from the traditional computational framework for trying to separate all the signals in the mixture input, which may fail to achieve a satisfactory performance in a complex auditory scene.

Several recent studies on human's auditory behavior and selective processing have revealed that the cortical activities of one listener in a multi-talker environment were dominated by the salient spectral and temporal features of the attended speaker, and were only weakly correlated with the unattended speaker (Mesgarani and Chang 2012; O'sullivan et al. 2015). Furthermore, studies on the neurophysiology of auditory selection showed that human's primary auditory cortex generates the frequency-selective attentional filter to tune into selective frequency channels (Da Costa et al. 2013) and the specific object representations in auditory memory enhances the perceptual precision of top-down attention (Lim, Wöstmann, and Obleser 2015). Based on the above observations, integrating selective attention and auditory memory into the computational auditory model would be a feasible solution for cocktail party problem.

Attention and Memory-Augmented Model

With the recent resurgence of interest of deep learning, many researchers have concentrated on using deep neural networks to map text, image and speech into a fixed-length vector (Sutskever, Vinyals, and Le 2014; Huang et al. 2014; LeCun, Bengio, and Hinton 2015). The main merit of these representation learning based methods is that they do not rely on any handcrafted features. However, the fixed-length vector representation is typically too small to accurately remember objects from the past, and may lose important details for response generation (Xu et al. 2016). To alleviate this drawback, lots of deep learning methods with explicit memory have been heavily studied recently, such as Memory Networks (MemNN) (Sukhbaatar et al. 2015), Neural Machine Translation (NMT) (Bahdanau, Cho, and Bengio 2015) and Neural Turing Machines (NTM) (Graves, Wayne, and Danihelka 2014). These methods exploit an external memory to store the input sequence with a continuous representation. Meanwhile, for automatically soft-searching the most related parts from the memory, they explore various attention mechanisms, such as $v^T h$ (dot), $v^T W h$ (general) and $g^T \tanh(W[v; h])$ (concat) (Luong, Pham, and Manning 2015), where v is the probe vector, h is the attended memory, and W and g are learned parameters.

However, the above memory modules work as a short-term memory (Chaudhuri and Fiete 2016) which usually persists the internal states during the current sample processing and resets the states when the next sample comes. In order to make the memory learn from the experience and recall prior knowledge explicitly, Kaiser et al. (Kaiser et al. 2017) give a novel long-term memory to accumulate the knowledge from training samples in a life-long manner. Along the direction of that work, we believe that introducing one life-long memory[1] into the computational auditory model can enhance the auditory selection performance, especially in top-down attention where the specific auditory object would be accumulated over time.

Our Approach

The goal of our approach is to solve top-down and bottom-up attention tasks in a unified computational auditory framework by modeling attention and memory. The main chal-

[1]It is worth noticing that we use the term "life-long memory" as mentioned in the previous work (Kaiser et al. 2017), and there is no difference with the term "long-term memory" in this work.

Table 1: Statistics of the selected datasets, including the mean length of the speech sounds, the male to female ratio, the number of train/dev/test set samples and the max supplementary stimuli for each unknown speaker (Supp (Spk)). Note that task 1 refers to the top-down attention, while task 2 refers to the bottom-up attention as described in Figure 1.

Datasets	Mean Len	Task 1 on Trained Speakers				Task 2 on Unknown Speakers		
		M:F	Train	Dev	Test	M:F	Test	Supp (Spk)
WSJ0	5.03 s	4:6	4,410	810	2,250	2:3	500	10
THCHS-30	9.15 s	2:8	4,410	810	2,250	1:4	500	10

lenges of this idea are that: (1) How does our method ASAM accumulate the prior knowledge into the life-long memory and use it to respond to the task-specific and stimulus-driven attention tasks; (2) How does the learned memory interact with the mixture input to attend and filter the target time-frequency units. The proposed method is described in detail in the following subsections.

Model Architecture

As described in Figure 2, given a raw mixture input x, our model first transforms it into time-frequency domain $X_{t,f}$ by Short-Time Fourier Transformation (STFT) and then maps it into d-dimensional embedding space $H_{t,f,d}$ via a Bidirectional Long-Short Term Memory (BiLSTM). In top-down attention scene, the target signal s is first fed into another BiLSTM and then accumulated into the life-long memory $\mathcal{M}$ corresponding with the attended speaker identity "Spk1" during training, and only the speaker identity "Spk1" is given to recall the prior knowledge during evaluation. While we directly use the model without retraining to test the bottom-up attention task with the speech stimulation $\hat{s}$ of the unknown speaker "Unk".

Accumulate the Prior Knowledge into Memory

To accumulate the speaker identity with their acoustic memory, one straightforward approach is to embed the i-th speaker identity p_i into d-dimensional vector by looking up a speaker embedding matrix $E \in \mathbb{R}^{|P|\times d}$, where E is a learned parameter and $|P|$ is the total number of speakers in the datasets. We denote this mode as ASAM-spk where the speaker embedding is served as a long-term memory as mentioned in (Kumar et al. 2016). Since this base model ASAM-spk directly trains the mapping matrix E rather than updates the memory based on the perceptual encoding of the salient speech, it may work well on top-down attention but could not be able to conduct the stimulus-driven attention task.

Inspired by (Kaiser et al. 2017), we put forward to accumulate the acoustic feature of the training samples into a life-long memory $\mathcal{M}$ following a speech perceptor, where the speech perceptor is implemented by a BiLSTM in our work and the memory $\mathcal{M}$ is formulated as a triple:

$$\mathcal{M} = (K_{memory}, V_{memory\times vector}, A_{memory}), \quad (1)$$

which consists of a vector K of memory keys, a matrix V of memory values and a vector A to track the age of items stored in memory. Given one training triple sample: (mixture input x, speaker identity p and target signal s), we first transform the target signal s into time-frequency domain $S_{t,f}$ by STFT, then apply an average pooling layer following BiLSTM on $S_{t,f}$ to extract the acoustic feature v. If the memory keyset already contains the given speaker identity p at the place n, then the age tracking value is reset and the value $V[n]$ is updated by taking the average of the current value and v and normalizing it:

$$A[n] \leftarrow 0,\ V[n] \leftarrow \frac{v + V[n]}{\|v + V[n]\|}. \quad (2)$$

Otherwise, we find a new or oldest place $n' = argmax_i A[i]$ in the memory and write the pair (p, v) there: $K[n'] \leftarrow p,\ V[n'] \leftarrow v,\ A[n'] \leftarrow 0$. With every memory update, the age tracking values of all non-updated indices also are incremented by 1.

In the test phase, our model can directly recall the corresponding memory vector v based on the given speaker identity p for task-specific task. While in stimulus-driven attention scene, the signal $\hat{s}$ is first encoded by the trained speech perceptor, BiLSTM, and then embed into the memory slot corresponding with "Unk" identity. In such way, the stimulus-driven task is unified with the task-specific procedure for the following continuous attention.

Obviously, if the memory size is larger than the total number of the speakers $|P|$, our model would memory all the acoustic features of the speakers. Otherwise, the forgetting mechanism would be triggered and the oldest memory would be erased. Specifically, we set the vector size to d and the memory size to $|P| + 1$ in our experiments. The extra memory is used to store the acoustic features of the unknown speakers temporarily, which usually occurs in the early stage of stimulus-driven task, and the temporal memory would be accumulated once the speaker identity is determined.

Be different from (Kaiser et al. 2017), we place the life-long memory at the encoder side for perception enhancement rather than the decoder side for response enhancement, which results in that the matrix representation in the memory is V rather than K. Considering the memory is trustworthy gradually in the life-long learning phase, we also update the memory with biased average rather than global average.

Attend the Target Time-Frequency Units

For the mixture input x, we first apply a BiLSTM layer along the time dimension of the mixture spectrogram X to compute the hidden state as follows:

$$h'_t = \overrightarrow{LSTM}(X_t) + \overleftarrow{LSTM}(X_t). \quad (3)$$

Table 2: Comparison of GNSDR results (mean±stdev) of our ASAM and baseline methods on the top-down attention task for two and three speaker mixture. DC (-40) means that the background noise threshold is set to -40 dB of the input's maximum magnitude, as well as DC (-60) and DC (-80) which ignore about 76.5%/82.6%, 41.5%/51.6% and 26.8%/32.2% time-frequency units on WSJ0/THCHS-30 respectively. ASAM-spk is similar to ASAM but the long-term memory is implemented with a speaker embedding matrix as described in Section . We randomly select a third speaker's speech and linearly mix it into the test dataset to form the three speaker mixtures (abbr. to Three). Two (noise) means that we mix some background noises (recorded in the street junction environment by (Barker et al. 2015)) into the two speaker mixtures in the test phase. (All of these models are trained on two speaker mixtures)

Methods	WSJ0			THCHS-30		
	Two	Three	Two (noise)	Two	Three	Two (noise)
DC	4.78±0.23	4.09±0.02	2.61±0.01	2.83±0.92	4.01±0.07	2.77±0.12
DC (-40)	7.47±0.07	**5.32±0.02**	3.29±0.04	6.56±0.08	5.48±0.15	2.81±0.23
DC (-60)	6.89±0.20	4.95±0.04	3.23±0.10	6.36±0.05	5.36±0.10	2.84±0.16
DC (-80)	6.82±0.05	4.94±0.12	3.74±0.26	5.76±0.33	4.82±0.17	3.49±0.22
ASAM-spk	**8.16±0.07**	5.06±0.07	3.92±0.16	**6.81±0.15**	**5.54±0.15**	**4.43±0.38**
ASAM	7.46±0.11	5.02±0.05	**4.36±0.13**	6.05±0.26	5.02±0.11	3.95±0.48

Then, the hidden state h'_t is fed into one feed-forward layer followed one reshape layer to generate the d-dimensional embedding vector $h_{t,f} \in \mathbb{R}^d$ of each time-frequency unit $X_{t,f}$. Finally, we perform selective attention on the mixture spectrogram X by using the probe vector v extracted from the life-long memory and the time-frequency hidden states $h_{t,f}$ as follows:

$$\alpha_{t,f} = sigmod(g^T \cdot \tanh(W \cdot v + U \cdot h_{t,f})), \quad (4)$$

where $g \in \mathbb{R}^{d\times 1}$, $W \in \mathbb{R}^{d\times d}$ and $U \in \mathbb{R}^{d\times d}$ are all learned parameters updated during training, and the attention weight $\alpha_{t,f}$ is adopted as frequency-selective attentional filter (Da Costa et al. 2013) or informational mask (Brungart 2001). Finally, our model ASAM produces the predicted spectrogram $X \times \alpha$, and minimizes the following objective function during training to learn the parameters:

$$\mathcal{L} = \sum_{t,f} \|S_{t,f} - X_{t,f} \times \alpha_{t,f}\|_2^2, \quad (5)$$

where S is the target spectrogram. Finally, the predicted signals are reconstructed by inverse STFT (iSTFT). From the above procedures, we can see that the probe vector v extracted from the life-long memory works as the attractor (magnet) generated from brain circuits (Kuhl 1991) to draw the attentive sounds.

Experiments

Below, a series of experiments are designed and conducted mainly to answer the questions: (1) How does the proposed approach ASAM compare with the state-of-the-art source separation method on top-down attention task; (2) Whether our approach can successfully attend a unknown target speaker on bottom-up attention task; (3) Whether a reasonable memory can be learned with only few samples of the target speaker.

Datasets and Setup

The auditory selection tasks are conducted on two selected datasets of speech mixtures based on the Wall Street Journal (WSJ0)[2] corpus (Garofalo et al. 2007) and Tsinghua Chinese 30 hour (THCHS-30)[3] database (Wang and Zhang 2015). We select 10 speakers with their part of speech sounds for top-down attention and other 5 speakers with their part of speech sounds for bottom-up attention from these datasets. We further linearly mix two speaker's utterances respectively. The mixtures generated from trained speakers are split into train, dev and test sets, while 10 clean utterances per unknown speaker are reserved as the stimuli. More summary statistics of these datasets are described in Table 1. From the statistics, we can see that the male to female ratio of THCHS-30 is biased which may make the task difficult.

In our experiments, the hyperparameters are set uniformly for the datasets and the baselines. All data are resampled to 8 kHz 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. In order to augment the variety of training samples, we circularly randomly shift (in the time domain) the signals and linearly mix two-speaker signals. We randomly generate 32 samples for one mini-batch, and per epoch contains 100 mini-batches. The learned parameters are all initialized randomly from a Glorot uniform distribution. Our models were trained using Nesterov Adam with a fixed learning rate of $\lambda = 0.002$. Training runs for up to 150 epochs with early stopping if the validation loss has not decreased for 10 epochs.

For architecture, we use a 2-layer BiLSTM with 300 hidden units in each direction for mixture encoder, and another 2-layer BiLSTM with 20 hidden units in each direction for speech perceptor along with the memory. The dimensions of T-F embeddings and memory vectors are all fixed to $d = 40$, resulting in a feed-forward layer of 5,160 hidden

[2]https://catalog.ldc.upenn.edu/LDC93S6A

[3]https://github.com/kaldi-asr/kaldi/tree/master/egs/thchs30

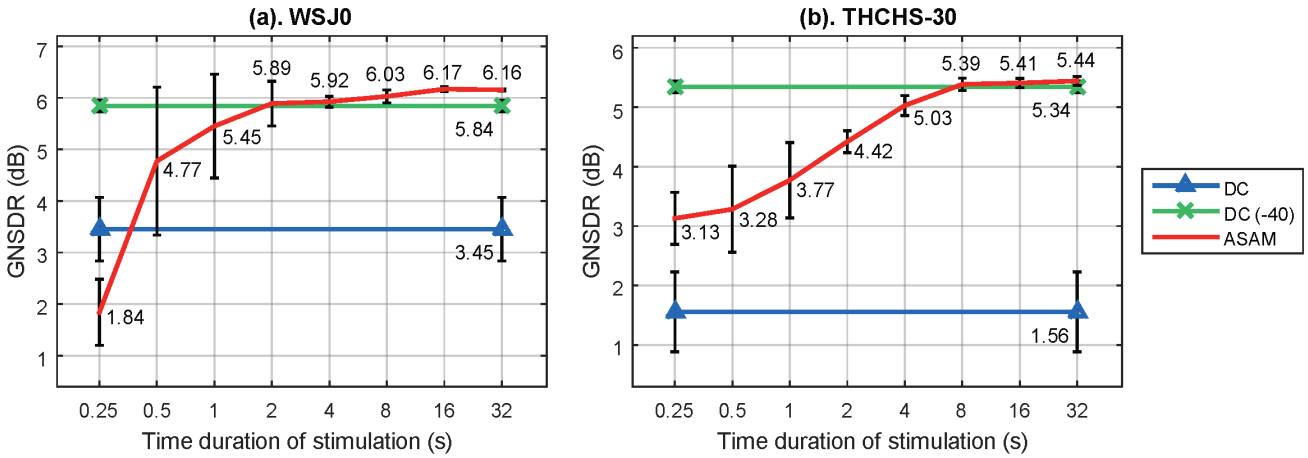

Figure 3: GNSDR results of our ASAM and baselines on the bottom-up attention task with two unknown speaker mixture and different time durations of the stimuli. For clarity, we only plot the base DC model and the best DC (-40) to compare our model, while the other models DC (-60) and DC (-80) get 5.23/5.17 and 4.98/4.56 NSDR results on WSJ0/THCHS-30 respectively.

units (40×129) after the mixture encoder. We constructed the baselines with the same configuration. To quantitatively evaluate the auditory selection results, we report the overall performance via the global signal-to-distortion improvement (GNSDR) using BSS_EVAL toolbox[4] (Vincent, Gribonval, and Févotte 2006). All experiments calculate the average results by repeating each experiment 5 times.

Results and Analysis

Performance on Top-down Attention In our experiments, the top-down auditory attention is to select the target speech based on one given speaker identity, which is a task-driven processing and usually needs to recall the prior knowledge. Table 2 shows the GNSDR results of our ASAM and other comparison methods which are all flexible to deal with two and three speaker mixtures, even though they are all trained on two speaker mixtures. We can see that ASAM-spk conducts best performances in most situations, and ASAM performs competitive results with the best DC based method DC (-40). However, the results of DC based methods is quite heavily influenced by the background noise threshold which is a tricky technology. Furthermore, DC based methods need to set the specific number of the expected clusters during evaluation. Since it is tough to determine the number of signal channels in the real auditory scene, DC based methods may fail in a complex auditory scene. In order to verify this assumption, we mix some real environment noises into two speaker mixture without retraining and report the results in Table 2. It is clear that ASAM based methods perform better robustness.

Performance on Bottom-up Attention Considering bottom-up auditory attention as stimulus-driven, we conduct the experiments on the unknown speaker mixtures. DC based methods directly cluster the mixture time-frequency units into different cluster spaces, while our ASAM uses the supplemental speech sounds as stimuli to attract the auditory attention. The GNSDR results are presented in Figure 3 and we can see that our ASAM outperforms the best DC based method after 2 s and 8 s respectively. Furthermore, as the time duration increases, the GNSDR performances become more stable with smaller deviation. Compared with the results in Table 2, the task-specific attention based on the accumulated prior knowledge indeed gains an advantage over the stimulus-driven attention based on the temporal acoustic features.

Effects of Attention with Different Amounts of Stimulus In order to make the effects of attention more intuitive, we give a visual example of auditory selective attention using our ASAM over one male and female mixture sample from WSJ0 test datasets in Figure 4. Figure 4(a) shows that the performances of our model on this sample are increasing as the time duration of the stimulus increasing, and Figure 4(f)-(g) present that our ASAM shifts attention toward the target speech by varying the time durations of the stimuli from 0.25 s to 1 s. The results reveal that the learned memory is sensitive to the amount of data and more samples per speaker can generate more reasonable memory. Despite this fact, another interesting result is that our model still generate a positive-going attentional filter (especially as shown near the right border of Figure 4(f)) based on few samples per speaker, such as the seeming chaotic stimulation with 0.25 s duration (as shown in Figure 4(b)).

Discussion

We do not do exhaustive searching for optimal architectures, but simply follow the previous work's parameter setting in DC (Hershey et al. 2016). Compared with DC based methods, our approach mainly have three advantages: (1) The attractor (magnet) is extracted from the life-long memory rather than the temporary input mixture, which would en-

[4]https://www.irisa.fr/metiss/bss_eval/

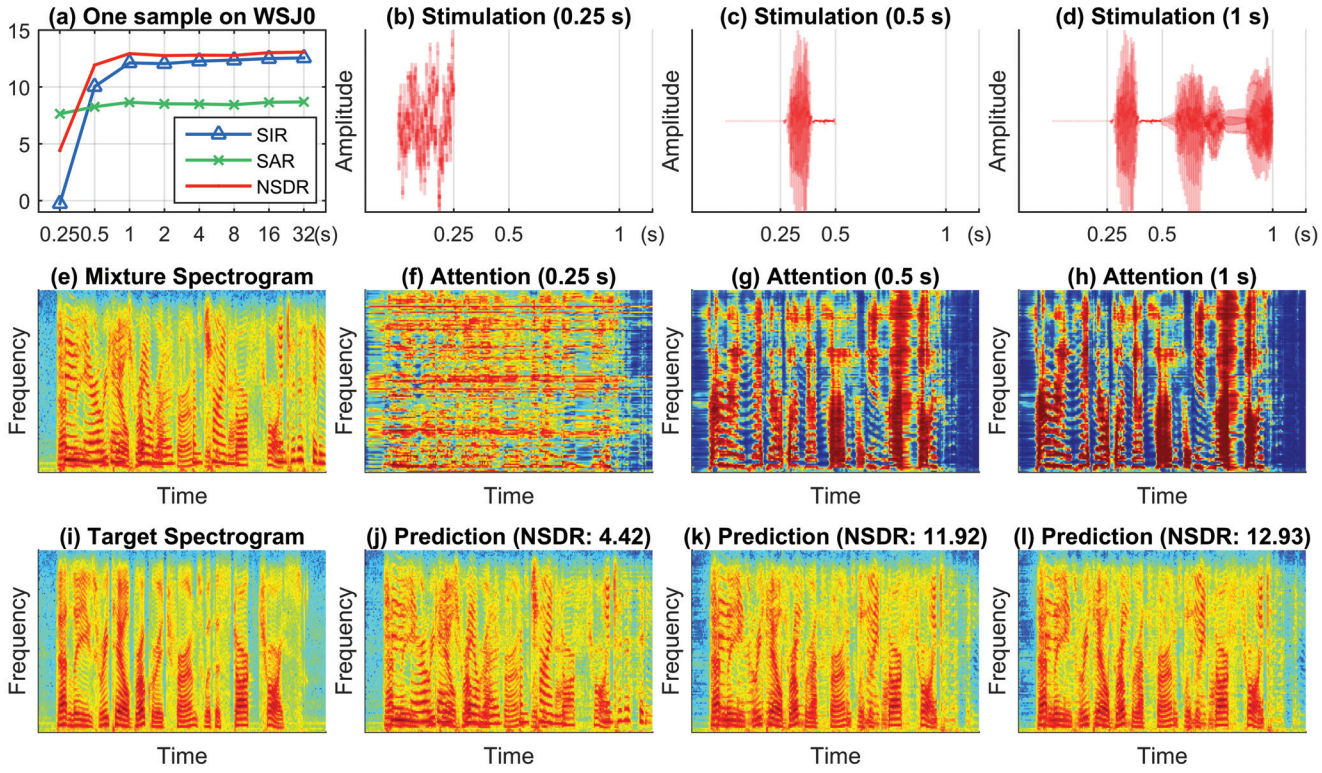

Figure 4: Effects of attention with different amounts of stimulus on one male and female mixture sample from WSJ0. (a) shows the SIR (Signal-to-Interference Ratio), SAR (Signal-to-Artifacts Ratio) and NSDR results, (b)-(d) are the auditory stimuli whose magnitudes are divided by the maximum magnitude, (e) is the mixture input spectrogram, (i) is the target spectrogram, (f)-(h) are attention maps based on the corresponding auditory stimuli and (j)-(l) are the corresponding predictions with their NSDR performances. (Best viewed in color)

sure that our approach uses the more reasonable attractor to attend the target speaker. (2) Our auditory attention module relaxes the strict assumption that the number of the input sources should be given during evaluation, which make our approach have the ability to deal with complex auditory scene. (3) Our auditory memory would be updated online once the speaker identity is determined after a salient speech stimulus, which helps our approach to enhance the learned memory online and even tackle the stimulus-driven attention on unknown speaker mixtures. Based on the above advantages, our approach performs competitive results on various experimental settings, such as closed/open speaker problems, two/three speaker inputs and mixing real background noises. An important issue worthy of research is how to evaluate cocktail party problem. Maybe conducting subjective listening and intelligibility test would be better, but yet Sources to Noise Ratio (SNR) based metrics, such as SDR, are objective and fair quantitative criteria, which do not rely on a particular type of separation algorithm, but simply try to compare the estimated signal with the target one (Kassebaum, Tenorio, and Schaefers 1990; Vincent, Gribonval, and Févotte 2006).

Conclusion and Future Works

In this work, we first defined two attention tasks for auditory selection: one is top-down task-specific attention to follow a conversation with a familiar friend and the other is bottom-up stimulus-driven attention to be attracted by a salient speech of a unknown speaker. Then we proposed a unified computational auditory model to solve the above two tasks based on the ideas from auditory selective attention of behavioral and cognitive neurosciences and from recent advances of memory-augmented neural networks. Finally, the experimental results on WSJ0 and THCHS-30 show that our model not only successfully conducts two attention tasks, but also performs robustness in a real noise environment.

This work attempts to bridge cognitive neurosciences and deep neural networks for designing a brain-inspired computational auditory model. We hope this solution would open up a new way for solving cocktail party problem. In future works, we intend to conduct further research on the switching and interaction between top-down and bottom-up attentions, or explore the potential of visual perception to improve auditory selection.

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB02070005), the National Natural Science Foundation (61602479) and the Independent Deployment Project of CAS Center for Excellence in Brain Science and Intelligent Technology (CEBSIT2017-02). We thank the reviewers for their insightful comments and also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

Barker, J.; Marxer, R.; Vincent, E.; and Watanabe, S. 2015. The third 'chime' speech separation and recognition challenge: Dataset, task and baselines. In *The 2015 IEEE Automatic Speech Recognition and Understanding Workshop*, 504–511.

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

Brungart, D. S. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* 109(3):1101–1109.

Chaudhuri, R., and Fiete, I. 2016. Computational principles of memory. *Nature neuroscience* 19(3):394–403.

Chen, Z.; Luo, Y.; and Mesgarani, N. 2017. Deep attractor network for single-microphone speaker separation. In *ICASSP*. IEEE.

Chen, Z. 2017. *Single Channel auditory source separation with neural network*. Ph.D. Dissertation, Columbia University.

Cherry, E. C. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25(5):975–979.

Da Costa, S.; van der Zwaag, W.; Miller, L. M.; Clarke, S.; and Saenz, M. 2013. Tuning in to sound: frequency-selective attentional filter in human primary auditory cortex. *Journal of Neuroscience* 33(5):1858–1863.

Garofalo, J.; Graff, D.; Paul, D.; and Pallett, D. 2007. Csr-i (wsj0) complete. *Linguistic Data Consortium, Philadelphia*.

Graves, A.; Wayne, G.; and Danihelka, I. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Hershey, J. R.; Chen, Z.; Le Roux, J.; and Watanabe, S. 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In *ICASSP*, 31–35. IEEE.

Huang, P.-S.; Kim, M.; Hasegawa-Johnson, M.; and Smaragdis, P. 2014. Deep learning for monaural speech separation. In *ICASSP*, 1562–1566. IEEE.

Kaiser, Ł.; Nachum, O.; Roy, A.; and Bengio, S. 2017. Learning to remember rare events. In *5th International Conference on Learning Representations*.

Kassebaum, J.; Tenorio, M. F.; and Schaefers, C. 1990. The cocktail party problem: speech/data signal separation comparison between backpropagation and sonn. In *Advances in Neural Information Processing Systems*, 542–549.

Kaya, E. M., and Elhilali, M. 2017. Modelling auditory attention. *Philosophical Transactions of the Royal Society of London* 372(1714).

Kuhl, P. K. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Attention Perception & Psychophysics* 50(2):93–107.

Kumar, A.; Irsoy, O.; Ondruska, P.; Iyyer, M.; Bradbury, J.; Gulrajani, I.; Zhong, V.; Paulus, R.; and Socher, R. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *The 33rd International Conference on Machine Learning*, 1378–1387.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436–444.

Lim, S.-J.; Wöstmann, M.; and Obleser, J. 2015. Selective attention to auditory memory neurally enhances perceptual precision. *Journal of Neuroscience* 35(49):16094–16104.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of Empirical Methods in Natural Language Processing*, 1412–1421.

Mesgarani, N., and Chang, E. F. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature* 485(7397):233–236.

O'sullivan, J. A.; Power, A. J.; Mesgarani, N.; Rajaram, S.; Foxe, J. J.; Shinn-Cunningham, B. G.; Slaney, M.; Shamma, S. A.; and Lalor, E. C. 2015. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral Cortex* 25(7):1697–1706.

Schmidt, M. N., and Olsson, R. K. 2006. Single-channel speech separation using sparse non-negative matrix factorization. In *INTERSPEECH*.

Sukhbaatar, S.; szlam, a.; Weston, J.; and Fergus, R. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, 2440–2448.

Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 3104–3112.

Vincent, E.; Gribonval, R.; and Févotte, C. 2006. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing* 14(4):1462–1469.

Wang, D., and Zhang, X. 2015. Thchs-30: A free chinese speech corpus. *arXiv preprint arXiv:1512.01882*.

Xu, J.; Shi, J.; Yao, Y.; Zheng, S.; and Xu, B. 2016. Hierarchical memory networks for answer selection on unknown words. In *26th International Conference on Computational Linguistics*, 2290–2299.

Yu, D.; Kolbæk, M.; Tan, Z.-H.; and Jensen, J. 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *ICASSP*.