BIO-MIMETIC ATTENTIONAL FEEDBACK IN MUSIC SOURCE SEPARATION

Ashwin Bellur, and Mounya Elhilali

Laboratory for Computational Audio Perception
Department of Electrical and Computer Engineering, Johns Hopkins University

ABSTRACT

Attention plays a vital role in helping us navigate our acoustic surroundings. It guides sensory processing to sift through the cacophony of sounds in everyday scenes and modulates the representation of targets sounds relative to distractors. While its conceptual role is well established, there are competing theories as to how attentional feedback operates in the brain and how its mechanistic underpinnings can be incorporated into computational systems. These interpretations differ in the manner in which attentional feedback operates as an information bottleneck to aid perception. One interpretation is that attention adapts the sensory mapping itself to encode only the target cues. An alternative interpretation is that attention behaves as a gain modulator that enhances the target cues after they are encoded. Further, the theory of temporal coherence states that attention seeks to bind temporally coherent features relative to anchor features as determined by prior knowledge of target objects. In this work, we study these competing theories within a deep-network framework for the task of music source separation. We show that these theories complement each other, and when employed together, yield state of the art performance in music source separation. We further show that systems with attentional mechanisms can be made to scale to mismatched conditions by retuning only the attentional modules with minimal data.

***Index Terms*—** Attention, bio-mimetic, music source separation, coherence, feature tuning

1. INTRODUCTION

Cognitive processes like attention play a significant role in our ability to navigate everyday acoustic environments. Attention essentially operates as an information bottleneck, enhancing acoustic cues of target sound objects while suppressing cues representing other competing objects in an auditory scene. This is borne out in several neurophysiological studies where selective attention was shown to emphasize the representation of target acoustic objects. Recordings of neural activity from individual neurons in primary auditory cortex of the mammalian brain have shown that attention to specific sounds induces rapid plasticity that adapts the tuning of these neurons in such a way that enhances the target and inhibits masker sounds [1, 2, 3]. Brain recordings from human listeners also shows that attending to a specific voice (particular speaker) in a noisy environment with competing speakers and reverberation induces enhanced encoding of the speech envelope of the target speaker [4, 5, 6, 7].

While the vital role of attention in accomplishing auditory tasks has been well established, there are several competing theories as to how these mechanisms can be interpreted and incorporated into computational audio processing systems. One interpretation posits that attention to a target acoustic object induces retuning of the sensory mapping that encodes the acoustic cues of the incoming signal in a manner that enhances the target sound relative to the maskers, as evidenced by rapid retuning of cortical neurons [8, 9]. Another interpretation postulates that attentional mechanisms operate at the perceptual stage, where attention performs a selection mechanism that modulates the output of the sensory mapping process, after the acoustic cues are already encoded. Recent work has proposed the concept of temporal coherence as means by which attention works with acoustic cues after they have been encoded to enhance perception of a desired object [10]. The principle of temporal coherence states that when attention is directed towards a cue of a target object, all features coherent with temporal activations of target cues become bound together such that the object of interest stands out [11, 12]. Thereby, attention biases the auditory system toward a particular grouping of encoded acoustic cues, depending on the attended object.

In this work, we explore a convolutional neural network (CNN) to leverage and appraise different interpretations of attention. This analysis is performed in the context of music source separation whereby the goal is to segregate different sources (vocals, instruments) from a single channel recording. The network is designed to take-in an input musical piece and output only the sources or acoustic objects towards which attention is being directed is retained. Using a network trained end-to-end for this particular task, we study the manifestation of different interpretations of attention. In particular, we explore two paradigms of attention: (1) using attention to retune the convolutional weights of the network hence shaping its selectivity; (2) using attention to modulate the output of

This work was supported by NIH U01AG058532 and R01HL133043, ONR N00014-19-1-2014, N00014-17-1-2736, and NSF 1734744.

the convolutional process using the principle of temporal coherence. Across these implementations, we examine the role of memory which represents the internal model of a target, which is then deployed to guide processing in the system. In other words, if one is attending to an acoustic object X, there is a presumption that one knows something about object X's characteristics and relies on their memory of that object to selectively attend to it. The current study explores specific implementations of this internal memory which is then deployed to facilitate attentional feedback.

It should be noted that the idea of attention has gained prominence in the deep learning literature across applications such as document classification [13], image captioning [14] and audio classification [15, 16, 17]. The manner in which attention operates across this body of work differs from the system in this study. Specifically, the machine learning attentional literature typically deploys attention as a soft-search mechanism for relevant words, pixels or audio events depending on task, while ignoring the interference. Attention effectively works as a gating operation to modulate embeddings of the neural network without explicit representations or memories of the target object. In contrast, the current work explores attention by explicitly training memories of target objects and evoking them during inference. A recent work in speech separation [18] attempted to explicitly train memories of speakers and recalling it to bias the network during inference with some success, similar to the proposed work. However, with the use of a long short term memory system for a very specific task, it is not clear how exactly attention is manifesting in aiding performance in this speech system.

The proposed system aims to contrast different interpretations of attention and examines their potentially complementary roles in aiding source separation. The paper presents a detailed scheme of different attention implementations within a CNN framework in section 2. In section 3, we provide details regarding the datasets used to validate and as well as the results, and comment on main conclusions in section 4.

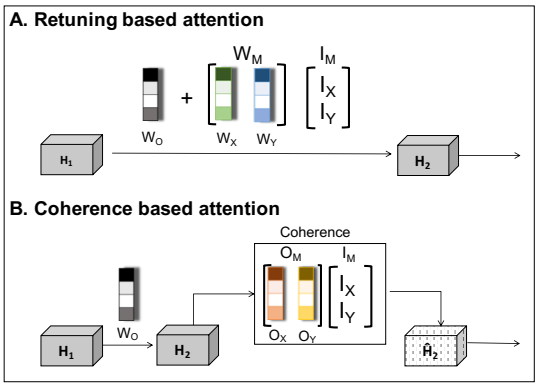

Fig. 1. Schematic for two attention paradigms. A.Retuning-based attention B.Coherence-based attention

2. DISTRIBUTED CNN WITH ATTENTION

In this work, we propose employing a convolutional neural network (CNN) equipped with attentional mechanisms deployed for the task of music source separation. The input to the CNN is the magnitude spectrogram of a piece of music as well as the identity of an acoustic object that the netwok intends to 'attend to'. The desired output will be the magnitude spectrogram of of only the acoustic object of interest while all other objects are suppressed. The output spectrogram of the network, along with the mixed phase of the musical piece is finally used to generate the waveform of the attended acoustic object. Before describing the overall architecture of the CNN, we look at the implementation of the two different interpretations of attention.

2.1. Retuning-based Attention

In the first implementation, we interpret attention as modulating the sensory mapping or re-tuning feature selectivity to enhance encoding of attended targets. In a CNN architecture, this implies that convolutional filters that map across layers are retuned by attentional feedback.

To achieve this outcome, we adopt a simple mechanism illustrated in figure 1A. Block H_1, of dimensions $m \times n \times o$, represents the embedding from a particular layer of the CNN. The dimensions m, n and o represent frequency channels, time frames and the number of hidden units respectively. W_O signifies a set of CNN filters of dimensions $p \times p \times o \times q$, representing the default mapping from embedding H_1 to H_2. The dimension p represents the height and width of the convolutional filter and q the number of hidden units in the embedding H_2. Let us assume that the dataset consists of two objects X and Y. Then W_X and W_Y in figure 1A, represent the desired retuning of W_O when attending to object X and Y respectively. I_X and I_Y are indicator variables that can take on binary values $\{0, 1\}$ indicating which object the network is directing its attention towards. If $I_X = 1$ and $I_Y = 1$, then network is attending to the music mixture and not just one of the acoustic objects.

In this interpretation of attention, W_X and W_Y can be viewed as a form of static *map memories* of acoustic objects X and Y which are employed whenever attention is invoked. The convolutional filters as well as the retuned weights for objects of interest are estimated during end-to-end training. This process can be succinctly represented as:

$$H_2 = f^{cnv}_{W_0 + W_M I_M}(H_1) \tag{1}$$

where f^{cnv} is the convolution operation, with W_M and I_M in the subscript representing the map memories and the indicator variables respectively. W_M is of dimensions $p \times p \times o \times q \times s$, where s is the number of acoustic objects in the dataset ($s = 2$) in the schematic shown in figure 1A.

2.2. Coherence-based Attention

In this paradigm, we employ the principle of temporal coherence to modulate the embeddings obtained after the convolution process as illustrated in figure 1B. The embeddings H_2 are obtained after convolving with the set of default CNN filters W_O; $H_2 = f^{cnv}_{W_0}(H_1)$. The embedding H_2 passes through the coherence block in figure 1B, to derive the modulated embedding $\hat{H}_2$.

In the coherence block, O_M denotes the anchor memory of dimensions $m \times q \times 2$. The symbols m and q as defined in the previous section, denote the frequency channels and number of hidden units in embedding H_2. O_X and O_Y in figure 1B, represent what we define as *anchor memories* of acoustic objects X and Y; similar to dictionary of basis in Non-negative matrix factorization techniques [19]. I_M of dimensions 2×1 contains the indicator variables.

The coherence block performs the following operation: Anchor $O_A = O_M * I_M$ is first estimated. It can be anchor memory of one of the objects or the sum of anchors of multiple objects, depending the values of the indicator variables given as input. Next, we determine the activation pattern R_A of the anchor of the attended object O_A.

$$R_A[1,t] = \sum_{i=1}^{m}\sum_{j=1}^{n} H_2[i,t,j] * O_A[i,j,1] \ \forall t \in \{1,...,n\} \tag{2}$$

we determine the modulated embeddings $\hat{H}_2$ using a non-linear operation:

$$\hat{H}_2 = H_2 \odot sigmoid(O_A * R_A) \tag{3}$$

where $\odot$ is the element-wise multiplication. The term $O_A * R_A$ determines the modulations for each of the dimensions of $\hat{H}_2$ as a product of the weights as represented in the anchor memory of the attended object and its activation at a particular time instant. The complete coherence operation is succinctly represented as:

$$\hat{H}_2 = f^{coh}_{O_M I_M}(H_2) \tag{4}$$

2.3. Distributed CNN

In order to test the proposed attentional mechanisms, we employ a CNN architecture shown in figure 2. The two interpretations of attentional mechanisms are leveraged at multiple vantage points of the CNN architecture. The CNN consists of 4 layers with leaky rectified linear unit (ReLU) activation and instead of a single hierarchy, there is branching in layer 2, with pooling in one of the sub-networks to capture slower spectrotemporal modulations. In order to estimate the magnitude spectrogram of the attended object, embeddings from lower layers are also used by concatenating the embeddings and using it as input for the final layer.

As indicated in figure 2, the convolutional filters are of dimensions 8×8 except for the final layer, with 2×2 max pooling to capture slower modulations at the branching off point in layer 2. To generate the network outputs, 1×1 convolution is performed in the final layer. The number of hidden units from layer 1 to layer 4 is 64, 256, 256 and 256 respectively. The network input is the magnitude spectrum of dimensions 512×64 , estimated using Short-Term Fourier Transform (STFT) with a window size of 1024 and hop size of 256. The music signal is downsampled to $8kHz$ before estimating the magnitude spectrogram. If X is the input spectrogram and Y the ground truth of the attended source or object, we employed the $L_{1,1}$ norm, $||Y - X||_{1,1}$ as the objective function to train the CNN. We shall train three CNN networks, one each with one of the interpretation of attentional mechanisms incorporated (just red or the blue blocks in figure 2) and one system with both of them.

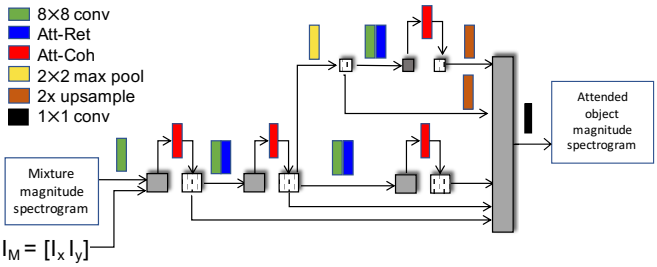

Fig. 2. Schematic of the distributed CNN system. Att-Ret refers to the retuning-based attention module and Att-Coh refers to the coherence-based attention module. I_M denotes to the indicator variables directing attention.

Such a distributed CNN architecture has basis in recent findings from functional magnetic resonance imaging (fMRI) studies of the human brain, where it was observed that acoustic cues are encoded at varying degrees of spectrotemporal resolutions [20, 21], using a spatially distributed neural network in the cortical regionsxz. This results in multiple redundant views of the input. These multiplexed views are hypothesized to enable segregation of acoustic objects and allow to discriminatively highlight distinct characteristics of objects of interest and distractors at multiple vantage points. The concept of employing embeddings at multiple levels of abstractions has also been found to extremely useful in multiple machine vision applications [22, 23] as well as audio applications like singing voice separation [24] , music source separation [25, 26, 27] and audio classification [16, 28].

3. EXPERIMENTS AND RESULTS

In order to validate the proposed system, we performed the task of music source separation on the DSD100 dataset [29]. The dataset consists of 100 songs, divided equally into 50 songs for training and 50 songs for testing. For each song, the ground truth consists of individual tracks from four sources (acoustic objects), bass, drums, vocal and other remaining in-

struments. In the context of this work, the network is given the mixed music as input as well as information regarding which of these 4 objects the networks is directing its attention towards. The desired output will be the magnitude spectrogram of just the objects for which the indicator variables (I_M) is equal to one. In the case where all the indicator variables are set to one, the systems behaves like an autoencoder, with desired output being the input spectrogram of the musical piece with all the objects present. The network was trained using Adam optimizer for 150,000 iterations with a learning rate of 10^{-4}. The training period was 33 hours on a single GPU.

Table 1 shows the results of music source separation in terms of signal to distortion (SDR) values, based on the BSS-EVAL metrics [30]. It can be seen that the proposed system with both retuning and coherence based attention *Att-Ret+Coh*, performs second best among the state of the art systems in all scenarios. It can also be seen that the *Att-Coh* (coherence based attention) performs better than *Att-Ret* (retuning based attention) across all conditions. This is understandable given that the modulation driven by coherence based attention depends on the stimulus and its coherence with the anchor memory. Whereas in the retuning based attention, the CNN filters are modulated with the same map memory of the attended object, irrespective of the specific input stimulus. It should also be noted that in the case of *MM-DenseNet*, the best performing system, four individual denoising style networks are trained for each of the sources, whereas *Hourglass* and *DeepNMF* are single networks, similar to the proposed work. However, it is significant that the proposed system (*Att-Ret+Coh*), a relatively shallow network compared to the other techniques, is able to perform on par with the state of the art systems with the attentional mechanisms incorporated.

Further, to show the ability of networks with attention to rapidly adapt in mismatched conditions, we also trained the proposed three networks using data from 2 performers, *abjones* and *amy*, from the MIR-1K database [31] consisting of Chinese karaoke songs. In this case, the dataset has just two acoustic objects vocals and accompaniments. The training period in this case, run for 15000, iterations was 3 hours. We then adapted the MIR-1K network for the DSD100 database by doing a single pass over the 50 training songs. During the adaptation, only the map memory and the anchor memory (blue and red blocks in figure 2) were retrained while keeping rest of the network fixed. To attend to the vocal sources of DSD100, we updated only the vocal memories intially trained using MIR-1K. In order to attend to the bass, drums and other sources, we individually retuned the accompaniments memory of MIR-1K to each of these three objects and then used the respective retuned memory during testing. This retuning of the MIR-1K memories using a single run of DSD100 train set, is accomplished in 2 hours. In order to compare the performance of the adapted systems, we also trained the Hourglass system using the MIR-1K data and retuned the complete network for the DSD100 dataset by similarly doing a single pass over the training set.

Table 2, shows the results on rapidly adapting the baseline and the proposed systems trained on the MIR-1K dataset using minimal training with DSD100 dataset. It can can seen that especially in the case of the *Att-Ret+Coh-Adap* system, which leverages both attentional mechanisms, the performance of the adapted system is on par with the other state of the art systems. For the bass class the *Att-Ret+Coh -Adap* performs better than fully trained systems like *DeepNMF* and *Hourglass*. For the remaining classes, the performance is approximately within 0.5 dB range of the fully trained systems. It can also be seen that the *Att-Ret+Coh-Adap* system performs much superior to the *Hourglass-Adap*, similarly trained with minimal data, for all classes with the exception of *others*. The hypothesis here is that the memories trained as part of the attention modules using MIR-1K have a reasonable baseline knowledge of the vocal and non-vocal objects in the space of music signals. In new conditions, just updating the memories which serve as informational bottlenecks, while keeping the underlying mapping fixed, can lead to versatile fast adapting systems retrained with minimal data.

Table 1. Median SDR values for music source separation on DSD100 dataset

Method	Bass	Drums	Others	Vocals
DeepNMF [32]	1.88	2.11	2.64	2.75
Hourglass [26]	1.77	4.11	2.36	5.16
MM-DenseNet [33]	3.91	5.37	3.81	6.00
Att-Ret	1.60	3.91	1.79	4.20
Att-Coh	2.01	4.23	2.04	4.62
Att-Ret+Coh	2.34	4.48	2.42	5.24

Table 2. Median SDR values for music source separation on DSD100 dataset on adapting the CNNs trained on MIR-1K

Method	Bass	Drums	Others	Vocals
Hourglass-Adap	1.65	2.70	1.90	3.92
Att-Ret-Adap	1.22	2.88	1.10	3.48
Att-Coh-Adap	1.46	2.56	1.08	3.75
Att-Ret+Coh-Adap	1.89	3.72	1.72	4.58

4. CONCLUSION

In this work, we incorporated two bio-mimetic interpretations of attention into a distributed CNN system, for the task of music source separation. We showed that combining the two interpretations enhances the performance of the system, as compared to just using one of the interpretations. We also highlighted a significant benefit of incorporating attentional mechanisms as information bottlenecks in a data driven system. We show that a network with attention capabilities trained on a particular database, can be rapidly adapted to scale to an altogether different mismatched database with minimal data, by retuning only the object memories in the attention modules.

5. REFERENCES

[1] K T Hill and L M Miller, "Auditory Attentional Control and Selection during Cocktail Party Listening," *Cerebral cortex (New York, N.Y.: 1991)*, vol. 20, no. 3, pp. 583–590, mar 2009.

[2] J. B. Fritz, M. Elhilali, and S. A. Shamma, "Adaptive Changes in Cortical Receptive Fields Induced by Attention to Complex Sounds," *Journal of Neurophysiology*, vol. 98, no. 4, pp. 2337–2346, 2007.

[3] P Yin, J B Fritz, and S A Shamma, "Rapid spectrotemporal plasticity in primary auditory cortex during behavior," *The Journal of neuroscience*, vol. 34, no. 12, pp. 4396–4408, mar 2014.

[4] 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, 2012.

[5] 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*, vol. 25, no. 7, pp. 1697–1706, jul 2015.

[6] Nai Ding and Jonathan Z. Simon, "Cortical entrainment to continuous speech: functional roles and interpretations," *Frontiers in Human Neuroscience*, vol. 8, may 2014.

[7] Søren Asp Fuglsang, Torsten Dau, and Jens Hjortkjær, "Noise-robust cortical tracking of attended speech in real-world acoustic scenes," *NeuroImage*, vol. 156, pp. 435–444, aug 2017.

[8] Jonathan B. Fritz, Mounya Elhilali, Stephen V. David, and Shihab A. Shamma, "Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in A1?," *Hearing research*, vol. 229, no. 1-2, pp. 186–203, jul 2007.

[9] S Shamma and J Fritz, "Adaptive auditory computations," *Current opinion in neurobiology*, vol. 25, pp. 164–168, apr 2014.

[10] Kai Lu, Yanbo Xu, Pingbo Yin, Andrew J Oxenham, Jonathan B Fritz, and Shihab A Shamma, "Temporal coherence structure rapidly shapes neuronal interactions," *Nature communications*, vol. 8, pp. 13900, 2017.

[11] Mounya Elhilali, Ling Ma, Christophe Micheyl, Andrew J. Oxenham, and Shihab A. Shamma, "Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes," *Neuron*, vol. 61, no. 2, pp. 317–329, jan 2009.

[12] Shihab A. Shamma, Mounya Elhilali, and Christophe Micheyl, "Temporal coherence and attention in auditory scene analysis," *Trends in neurosciences*, vol. 34, no. 3, pp. 114–23, mar 2011.

[13] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016, pp. 1480–1489.

[14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[15] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 121–125.

[16] Changsong Yu, Karim Said Barsim, Qiuqiang Kong, and Bin Yang, "Multi-level Attention Model for Weakly Supervised Audio Classification," *arXiv preprint arXiv:1803.02353*, 2018.

[17] Qiuqiang Kong, Changsong Yu, Turab Iqbal, Yong Xu, Wenwu Wang, and Mark D Plumbley, "Weakly labelled audioset classification with attention neural networks," *arXiv preprint arXiv:1903.00765*, 2019.

[18] Jiaming Xu, Jing Shi, Guangcan Liu, Xiuyi Chen, and Bo Xu, "Modeling attention and memory for auditory selection in a cocktail party environment," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[19] Paris Smaragdis and Judith C Brown, "Non-negative matrix factorization for polyphonic music transcription," 2003.

[20] Roberta Santoro, Michelle Moerel, Federico De Martino, Rainer Goebel, Kamil Ugurbil, Essa Yacoub, and Elia Formisano, "Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex," *PLoS Computational Biology*, vol. 10, no. 1, 2014.

[21] Roberta Santoro, Michelle Moerel, Federico De Martino, Giancarlo Valente, Kamil Ugurbil, Essa Yacoub, and Elia Formisano, "Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 114, no. 18, pp. 4799–4804, may 2017.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[24] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde, "Singing voice separation with deep u-net convolutional networks," 2017.

[25] Jen-Yu Liu and Yi-Hsuan Yang, "Denoising auto-encoder with recurrent skip connections and residual regression for music source separation," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 773–778.

[26] Sungheon Park, Taehoon Kim, Kyogu Lee, and Nojun Kwak, "Music source separation using stacked hourglass networks," *arXiv preprint arXiv:1805.08559*, 2018.

[27] Emad M Grais, Hagen Wierstorf, Dominic Ward, and Mark D Plumbley, "Multi-resolution fully convolutional neural networks for monaural audio source separation," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 340–350.

[28] Ashwin Bellur and Mounya Elhilali, "Audio object classification using distributed beliefs and attention," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. PP, pp. 1–1, 01 2020.

[29] Fabian-Robert Stöter, Antoine Liutkus, and Nobutaka Ito, "The 2018 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2018, pp. 293–305.

[30] Emmanuel Vincent, R 'emi Gribonval, and C 'edric F 'evotte, "Performance Measurement in Blind Audio Source Separation," vol. 14, no. 4, pp. 1462, 2006.

[31] Chao-Ling Hsu and Jyh-Shing Roger Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, 2009.

[32] J Le Roux, J R Hershey, and F Weninger, "Deep NMF for speech separation," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 66–70.

[33] Naoya Takahashi and Yuki Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 21–25.