Joint Localization and Classification of Multiple Sound Sources Using a Multi-task Neural Network

Weipeng He[1,2]*, Petr Motlicek*[1] *and Jean-Marc Odobez*[1,2]

[1]Idiap Research Institute, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{weipeng.he, petr.motlicek, odobez}@idiap.ch

Abstract

We propose a novel multi-task neural network-based approach for joint sound source localization and speech/non-speech classification in noisy environments. The network takes raw short time Fourier transform as input and outputs the likelihood values for the two tasks, which are used for the simultaneous detection, localization and classification of an unknown number of overlapping sound sources, Tested with real recorded data, our method achieves significantly better performance in terms of speech/non-speech classification and localization of speech sources, compared to method that performs localization and classification separately. In addition, we demonstrate that incorporating the temporal context can further improve the performance.

Index Terms: sound source localization, speech/non-speech classification, computational auditory scene analysis, deep neural network, multi-task learning

1. Introduction

Sound source localization (SSL) is essential to many applications such as perception in human-robot interaction (HRI), speaker tracking in teleconferencing, etc. Precise localization of sound sources provides the prerequisite information for speech/signal enhancement, as well as subsequent speaker identification, automatic speech recognition and sound event detection. Although many approaches have addressed the problem of SSL, there have been only a few studies on the discrimination of the interfering noise sources from the target speech sources in noisy environments.

Traditional signal processing-based sound source localization methods [1–3] rely heavily on ideal assumptions, such as that the noise is white, the SNR is greater than 0dB, the number of sources is known, etc. However, in many real HRI scenarios (e.g. HRI in public places [4]), where the environment is wild and noisy, the aforementioned assumptions hardly hold. We aim to develop SSL methods under the following challenging conditions:

(C1) An unknown number of simultaneous sound sources.

(C2) Presence of strong robot ego-noise.

(C3) Presence of directional interfering non-speech sources in addition to the speech sources.

It has been shown recently that the deep neural networks-based (DNN) approaches significantly outperform traditional signal processing-based methods in localizing multiple sound sources under the conditions (C1) and (C2) [5]. The DNN approaches directly learn to approximate the unknown and complicated mapping from input features to the directions of arrival (DOAs) from a large amount of data without making strong assumption about the environment. In addition, the spectral characteristics of the robot ego-noise can be implicitly learned by the neural networks. However, under condition (C3), this approach does not discriminate the noise sources from the speech sources, and we have observed that this method is sensitive to non-speech sound sources, for instance keyboard clicking, crumpling paper, and footsteps, all of which produce false alarms.

Sound source localization in the presence of interfering noise sources has been studied by applying classification on sources from individual directions [6, 7]. In contrast to conventional speech/non-speech (SNS) classification problem, which takes a one-channel signal as input, the sound classification of multiple signals needs to extract the source signal from the mixed audio prior to applying classification. The methods for extraction include beamforming [7] and sound source separation by time-frequency masking [6]. Both methods apply disjoint source localization and classification. Specifically, the classification is either independent or subsequent of the localization.

Localization and classification of sources in sound mixtures are closely related. The localization helps the classification by providing spatial information for better separation or enhancement of sources. Vice versa, knowing the types of the sources provides the spectral information that helps the localization. However, there has been little discussion on simultaneous localization and classification of sound sources.

In this paper, we address how to solve source localization and classification jointly in noisy HRI scenarios by a deep multi-task neural network.

2. Approach

We propose a deep convolutional neural network with multi-task outputs for the joint localization and classification of sources (Fig. 2). In the rest of this section, we introduce the network input/output, loss functions, network architectures and its extension by taking temporal context as input.

2.1. Network Input

We adopt the raw short time Fourier transform (STFT) as the input, as it contains all the required information for both tasks. This contrasts with previous works, in which the features for these two tasks are radically different. Sound source localization relies on the inter-channel features (e.g. cross-correlation [1, 5, 8], inter-channel phase and level difference [9, 10]) or the subspace-based features [2, 11, 12], whereas SNS classification normally requires features computed from the power spectrum [13, 14]. Recently, it has been shown that

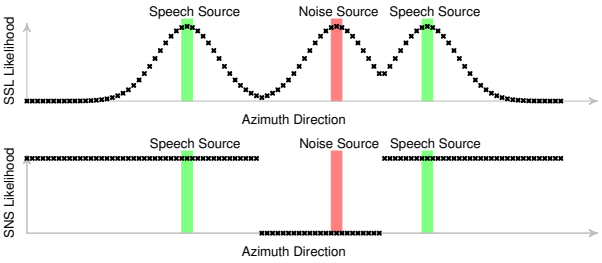

Figure 1: *Desired output of the multi-task network.*

instead of applying complicated feature extraction, we can directly use the power spectrum as the inputs for neural network-based sound source localization [15]. However, unlike in [15], our method employs the real and imaginary parts of the STFT, preserving both the power and phase information.

The raw data received by the robot are 4-channel audio signals sampled at 48 kHz. Their STFT is computed in frames of 2048 samples (43 ms) with 50% overlap. Then, a block of 7 consecutive frames (170 ms) are considered a unit for analysis. The 337 frequency bins between 100 and 8000 Hz are used. The real and imaginary parts of the STFT coefficients are split into two individual channels. Therefore, the result input feature of each unit has a dimension of $7 \times 337 \times 8$ (temporal frames $\times$ frequency bins $\times$ channels).

2.2. Network Output and Loss Function

The multi-task network outputs on each direction, the likelihood of the presence of a sound source, $\mathbf{p} = \{p_i\}$, and the likelihood of the sound being a speech source, $\mathbf{q} = \{q_i\}$. The elements p_i and q_i are associated with one of the 360 azimuth directions θ_i.

Based on the likelihood-based coding in [5], the desired SSL output values are the maximum of Gaussian functions centered at the DOAs of the ground truth sources (Fig 1):

$$p_i = \begin{cases} \max_{\bar{\theta} \in \Theta} \left\{ e^{-d(\theta_i, \bar{\theta})^2 / \sigma^2} \right\} & \text{if } |\Theta| > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

where $\Theta = \Theta^{(s)} \cup \Theta^{(n)}$ is the union of the ground truth speech source and interfering source DOAs, σ is the parameter to control the width of the Gaussian curves, $d(\cdot, \cdot)$ denotes the azimuth angular distance, and $|\cdot|$ denotes the cardinality of a set.

The desired SNS output values are either 1 or 0 depending on the type of the nearest source[1] (Fig 1):

$$q_i = \begin{cases} 1 & \text{if the nearest source is speech} \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

Loss function. The loss function is defined as the sum of the mean squared error (MSE) of both predictions:

$$\text{Loss} = \|\hat{\mathbf{p}} - \mathbf{p}\|_2^2 + \mu \sum_i w_i \left| \hat{q}_i - q_i \right|^2, \quad (3)$$

where $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ are the network outputs, $\mathbf{p}$ and $\mathbf{q}$ are the desired outputs, and μ is a constant. The SNS loss is weighted by $\{w_i\}$, which depends on its distance to the nearest source (w_i differs from p_i only in the parameter for curve width σ_w):

$$w_i = \begin{cases} \max_{\bar{\theta} \in \Theta} \left\{ e^{-d(\theta_i, \bar{\theta})^2 / \sigma_w^2} \right\} & \text{if } |\Theta| > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

[1] It is assumed that sources are not co-located.

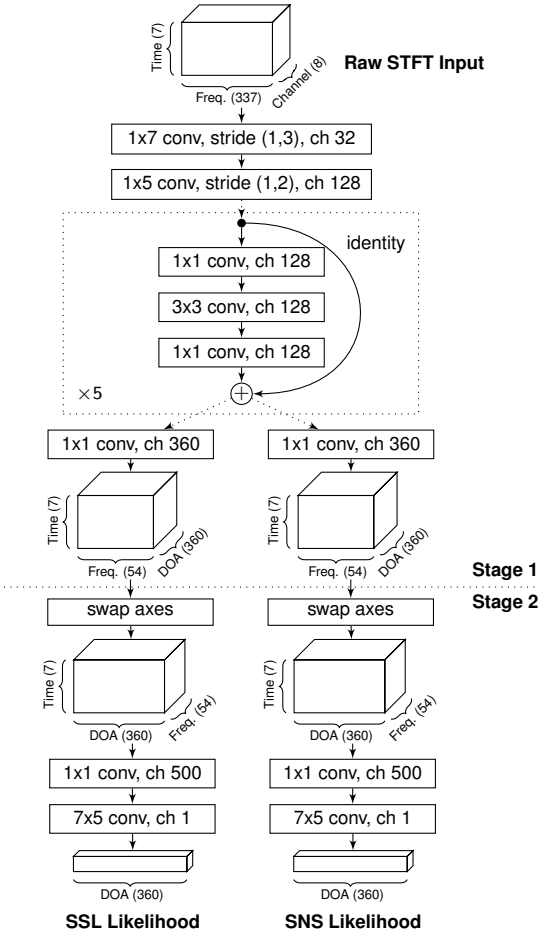

Figure 2: *The architecture of the multi-task network.*

so that the network is trained with the emphasis around the directions of the active sources.

Decoding. During test, the method localizes the sound sources by finding the peaks in the SSL likelihood that are above a given threshold:

$$\hat{\Theta} = \left\{ \theta_i : p_i > \xi \quad \text{and} \quad p_i = \max_{d(\theta_j, \theta_i) < \sigma_n} p_j \right\}, \quad (5)$$

where ξ is the prediction threshold and σ_n is the neighborhood distance for peak finding. Furthermore, to predict the DOAs of speech sources, we combine the SSL and SNS likelihood to further refine the peaks in the SSL likelihood:

$$\hat{\Theta}^{(s)} = \left\{ \theta_i : p_i q_i > \xi \quad \text{and} \quad p_i = \max_{d(\theta_j, \theta_i) < \sigma_n} p_j \right\}. \quad (6)$$

We set $\sigma = \sigma_n = 8°$, $\mu = 1$ and $\sigma_w = 16°$ in the experiments.

2.3. Network Architecture

The multi-task network is a fully convolutional neural network consisting of a residual network (ResNet [16]) common trunk and two task-specific branches (Fig. 2). The common trunk starts with the reduction of the size in the frequency dimension by two layers of strided convolution. These initial layers are followed by five residual blocks. The identity mappings in the residual blocks allow a deeper network to be trained without being affected by the vanishing gradients problem. It has

been shown that the ResNet is effective for sound source localization problem [15]. The hard parameter sharing in such common trunk provides regularization and reduces the risk of overfitting [17].

The task-specific branches are identical in structure. They both start with a 1×1 convolutional layer with 360 output channels (corresponding to 360 azimuth directions). The layers until this point represent *Stage 1*, in which all the convolutions are along the time-frequency (TF) domain, therefore the outputs have local receptive fields in the TF domain and can be considered as the initial estimation (of SSL and SNS) for individual TF points. In the rest of the network, *Stage 2*, the convolutions are local in time and DOA dimensions but global in the frequency dimension. Technically, this is achieved by swapping the DOA and the frequency axes. The final output of each branch is a 360-dimension vector indicating the likelihood of SSL and SNS respectively. In addition, the batch normalization [18] and rectified linear unit (ReLU) activation function [19] are applied between all convolutional layers.

2.4. Two-Stage Training

We train the network from scratch with a two-stage training scheme inspired by [5]. We first train *Stage 1* for four epochs by imposing supervision to its output. The loss function at this stage is defined as the sum of Eq. 3 applied to all the TF points[2]. Such supervision provides a better initialization of the *Stage 1* parameters for further training.

Then, the whole network is trained in an end-to-end fashion (using the loss function of Eq. 3 at the end) for ten epochs. We use the Adam optimizer [20] with mini-batches of size 128 for training.

2.5. Adding Temporal Context

The multi-task network can be simply extended to incorporate the temporal context to the input. That is, in addition to the block of 7 frames to be analyzed (i.e. for which we want to make a prediction), we add 10 frames (210 ms) in the past and 10 frames (210 ms) in the future as input to the network, thus reaching an input duration of 600 ms. As the network is fully convolutional, its structure remains the same except for the last convolutional layer where the kernel shape is changed from 7×5 to 27×5 (temporal frames $\times$ DOA).

3. Experiments

We collected noisy recordings with our robot Pepper, which has four coplanar microphones on its head[3], and evaluated the performance of the methods in terms of sound localization, SNS classification, as well as speech localization.

3.1. Data

The collected recordings consist of two sets: the loudspeaker mixtures and human recordings (Table 1). The loudspeaker mixture recordings are an extension of the loudspeaker dataset from [5] by mixing new non-speech recordings with the speech recordings. The non-speech recordings were collected by playing non-speech audio segments from loudspeakers in the same condition as the speech recordings. These segments are from the Audio Set [21] and cover a wide range of audio classes, including a variety of noises, music, singing, non-verbal human sounds, etc.

Table 1: *Specifications of the recorded data.* 360° *means the source can be from any azimuth direction.* FoV *is the camera's field of view.*

	Loudspeaker		Human
	Training	Test	Test
Total duration	32 hours	17 hours	8 min
Max. # of speech	2	2	3
Max. # of noise	1	1	1
# of speakers	148	16	7
DOA range (speech)	360°	360°	in FoV
DOA range (noise)	360°	360°	360°

The human recordings involve people having natural conversation or reading with provided scripts while non-speech segments were played from loudspeakers. Ground truth source locations were automatically annotated and the voice activity detection was manually labelled.

3.2. Methods for Comparison

We include the following methods for comparison:

MTNN The proposed multi-task network.

MTNN-CTX The proposed multi-task network with temporal context extension.

MTNN-N2S The proposed multi-task network trained without the two-stage scheme.

SSLNN A single-task network (same structure as in Fig. 2 but only with one output branch) for sound localization.

SpeechNN A single-task network for speech localization (trained to only localize speech sources).

SSL+BF+SNS It first localizes sounds with the SSLNN, then extracts the signals from the candidate DOAs by the minimum variance distortionless response (MVDR) beamformer [22], and finally classifies their sound type with a SNS neural network (similar ResNet structure).

SRP-PHAT steered response power with phase transform [3].

3.3. Sound Source Localization Results

We evaluate the sound source localization as a detection problem, where the number of sources is not a priori known. To do this, we compute the precision and recall with a varying prediction threshold ξ of Eq. 5. A prediction is considered to be correct if it is within 5° of error from a ground truth DOA. Then, we plot the precision vs. recall curves on the two datasets (a) loudspeaker mixtures (b) human recordings (Fig. 3). The proposed multitask network achieves more than 90% accuracy and 80% recall on both datasets, and is only slightly worse than the single-task network trained for sound source localization. Note that all neural network-based methods are significantly better than SRP-PHAT.

3.4. Speech/Non-Speech Classification Results

To evaluate the performance of speech/non-speech classification, we compute the classification accuracy under two conditions: considering the SNS predictions (1) in the ground truth directions, and (2) in the predicted directions (Table 2). Specifically, under condition (1), for each ground truth sound source,

[2]We don't use individual ground truth for each TF point, because it is impractical to acquire.

[3]`http://doc.aldebaran.com/2-5/family/pepper_technical/microphone_pep.html`

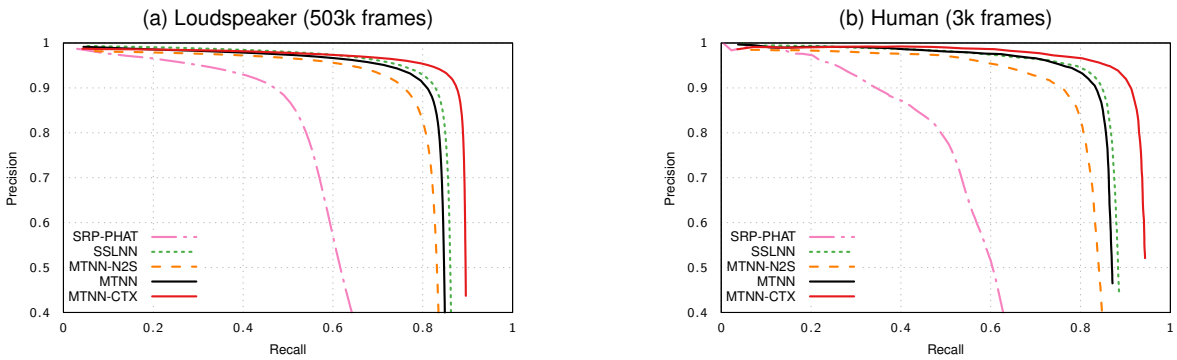

Figure 3: *Sound source localization performance.*

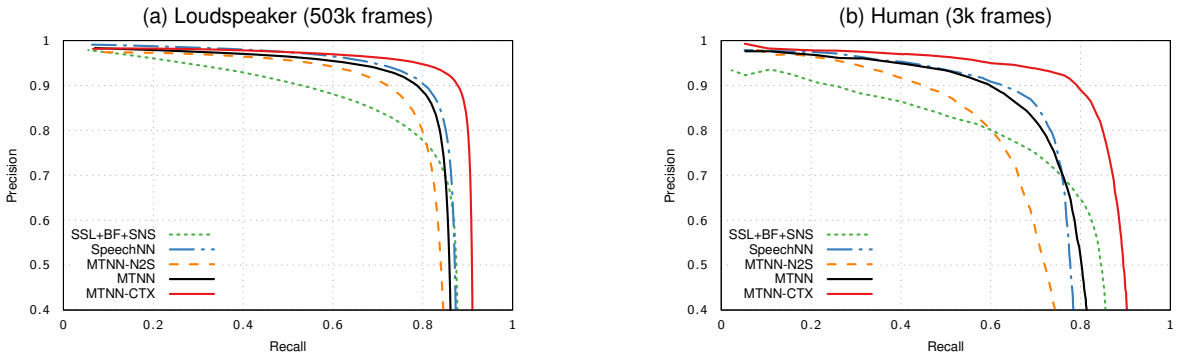

Figure 4: *Speech source localization performance.*

Table 2: *Speech/non-Speech classification accuracy. Numbers in the parentheses indicate the recall of the DOA prediction.*

Dataset	Loudspeaker		Human	
Directions	G.T.	Pred. (Rec.)	G.T.	Pred. (Rec.)
SSL+BF+SNS	0.80	0.81 (0.83)	0.68	0.73 (0.83)
MTNN-N2S	0.93	0.96 (0.79)	0.82	0.83 (0.76)
MTNN	0.95	0.97 (0.81)	0.85	0.86 (0.82)
MTNN-CTX	0.96	0.98 (0.85)	0.89	0.89 (0.86)

we check how accurate the method predict its type in the ground truth DOA. Such evaluation is independent of the localization method. Under condition (2), we first select the predicted DOAs that are close to the ground truth (error $< 5°$), and then evaluate the SNS accuracy on these directions. In this case, not all ground truth sources are matched to a prediction (recall < 1) and the result is dependent on the localization method. This is why the performance in the predicted DOAs can be better than that in the ground truth DOAs. We make the DOA prediction by Eq. 5 with $\xi = 0.5$.

Our proposed method achieves more than 95% of accuracy in the loudspeaker recordings and more than 85% accuracy in the human recordings. All the multi-task approaches are significantly better than SSL+BF+SNS, which extracts signal by beamforming and then classifies.

3.5. Speech Source Localization Results

We evaluated the speech source localization performance in the same way as that for sound source localization (Fig. 4). In terms of speech localization, the multi-task approaches significantly outperform the SSL+BF+SNS, due to their better performance in classification. The proposed method is slightly worse than the single-task network for speech localization in the loudspeaker recordings, and achieves similar performance in the human recordings.

3.6. Two-stage Training and Temporal Context

In all the three tasks, the proposed method trained in two stages is superior than the one trained with only the end-to-end stage. This implies that the two-stage training scheme effectively helps the training process.

In addition, we see that adding temporal context improves both the sound source localization and classification performance, and as a result, greatly improves the speech localization performance. Demonstration videos of the proposed method are available in the supplementary material.

4. Conclusion

In this paper, we have described of a novel multi-task neural network approach for joint sound source localization and speech/non-speech classification. The proposed method achieves significantly better results in term of speech/non-speech classification and speech source localization, compared to method that separates localization and classification. We further improve the performance with a simple extension of the method by adding temporal context to inputs.

5. Acknowledgements

This research has been partially funded by the European Commission Horizon 2020 Research and Innovation Program under grant agreement no. 688147 (MultiModal Mall Entertainment Robot, MuMMER, mummer-project.eu).

6. References

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[3] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Apr. 1997, pp. 375–378 vol.1.

[4] M. E. Foster, R. Alami, O. Gestranius, O. Lemon, M. Niemel, J.-M. Odobez, and A. K. Pandey, "The MuMMER Project: Engaging Human-Robot Interaction in Real-World Public Spaces," in *Social Robotics*. Springer, Cham, Nov. 2016, pp. 753–763.

[5] W. He, P. Motlicek, and J.-M. Odobez, "Deep Neural Networks for Multiple Speaker Detection and Localization," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, May 2018.

[6] T. May, S. v. d. Par, and A. Kohlrausch, "A Binaural Scene Analyzer for Joint Localization and Recognition of Speakers in the Presence of Interfering Noise Sources and Reverberation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2016–2030, Sep. 2012.

[7] M. Crocco, S. Martelli, A. Trucco, A. Zunino, and V. Murino, "Audio Tracking in Noisy Environments by Acoustic Map and Spectral Signature," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–14, 2017.

[8] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 5745–5749.

[9] M. S. Datum, F. Palmieri, and A. Moiseff, "An artificial neural network for sound localization using binaural cues," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 372–383, Jul. 1996.

[10] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," *Proceedings of Interspeech 2015*, pp. 3302–3306, 2015.

[11] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 405–409.

[12] ——, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2016, pp. 603–609.

[13] A. Martin, D. Charlet, and L. Mauuary, "Robust speech/non-speech detection using LDA applied to MFCC," in *2001 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 2001, pp. 237–240 vol.1.

[14] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 7378–7382.

[15] N. Yalta, K. Nakadai, and T. Ogata, "Sound Source Localization Using Deep Learning Models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, Feb. 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.

[17] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv:1706.05098 [cs, stat]*, Jun. 2017, arXiv: 1706.05098. [Online]. Available: http://arxiv.org/abs/1706.05098

[18] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *PMLR*, Jun. 2015, pp. 448–456.

[19] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[20] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Dec. 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[21] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[22] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, Oct. 1987.