ROBUST DIRECTION ESTIMATION WITH CONVOLUTIONAL NEURAL NETWORKS BASED STEERED RESPONSE POWER

Pasi Pertilä, Emre Cakir

Department of Signal Processing, Tampere University of Technology, Finland

ABSTRACT

The steered response power (SRP) methods can be used to build a map of sound direction likelihood. In the presence of interference and reverberation, the map will exhibit multiple peaks with heights related to the corresponding sound's spectral content. Often in realistic use cases, the target of interest (such as speech) can exhibit a lower peak compared to an interference source. This will corrupt any direction dependent method, such as beamforming.

Regression has been used to predict time-frequency (TF) regions corrupted by reverberation, and static broadband noise can be efficiently estimated for TF points. TF regions dominated by noise or reverberation can then be de-emphasized to obtain more reliable source direction estimates. In this work, we propose the use of convolutional neural networks (CNNs) for the prediction of a TF mask for emphasizing the direct path speech signal in time-varying interference. SRP with phase transform (SRP-PHAT) combined with the CNN-based masking is shown to be capable of reducing the impact of time-varying interference for speaker direction estimation using real speech sources in reverberation.

Index Terms— sound source localization, steered response power, convolutional neural networks, time-frequency masking

1. INTRODUCTION

The traditional methods for localization of a sound source using a microphone array exploit the observed wavefront's time differences of arrival (TDOA). With the knowledge of microphone positions, propagation speed of sound, and estimates of TDOA values between microphone pairs, the source position can be estimated using a closed-form solution or an iterative approach, refer to [1] for a review of such methods. The generalized cross-correlation (GCC) is a popular method for TDOA extraction, and GCC with phase transform (GCC-PHAT) has been widely used due to its robustness. However, the successful extraction of TDOA values depends on the SNR, the level of interfering sources, and the amount of reverberation. Reverberation even in moderate amounts is able to deteriorate TDOA estimation [2]. In [3] ridge-regression is used to learn the localization precision of a TF point, that is used to weight the GCC-PHAT to improve TDOA estimation in a simulated setting. GCC-PHAT can be also enhanced with a TF mask that aims to remove components of broadband noise. In this direction a continuous [4] and binary weighted TF mask based on SNR estimation is proposed in [5], which was shown to lessen the effects of coherent broadband noise in TDOA estimation.

The SRP method is a more robust localization approach than the TDOA approach. Instead of extracting a single TDOA value for each microphone pair, the array is steered to different directions. The direction with the most power represents the dominant source direction. SRP based on the GCC-PHAT (SRP-PHAT) removes the amplitude information by whitening the signal and thus utilizes only the cross-spectrum phase information. It is typically adopted in indoor localization [6].

Binary TF masks have been applied in computational auditory scene analysis (CASA) to separate the components of a spectrogram caused by different sound sources [7]. Separation is performed by multiplying the noisy spectrogram with the mask. Continuous real-valued masks have been found less susceptible to musical artifacts. The classical Wiener-filter is an example of a continuous TF mask. Recent deep neural network (DNN)-based single channel speech enhancement methods have been successful in learning to predict TF masks in various noisy conditions [8, 9].

Convolutional neural networks (CNNs) are discriminative classifiers that compute their neuron activations through shared weights over local receptive fields. They have been used widely and produced state-of-the-art results in classification tasks such as image recognition [10], speech recognition [11] and acoustic event detection [12]. Apart from classification, CNNs have also been proposed for regression tasks where both input and target output consists of multi-dimensional data such as images. Image restoration/denoising [13], image super-resolution [14] and weighted mask estimation for speech source separation [15] can be listed as examples of CNN applied over regression tasks.

This paper is inspired by the success of DNN-based TF mask learning for speech enhancement and considers the approach in speaker localization. In contrast to previous TF mask-based TDOA estimators using GCC-PHAT in the presence of static noise and/or reverberation, we consider the SRP-PHAT approach in presence of reverberation with non-stationary interference. Speech signals contain significant local information in spectral domain. However, the spectral position of this information may exhibit some shift due to changing speaker and environmental conditions. The translational shift invariance property of CNNs make them a suitable option for our task, and therefore we propose to use CNNs for TF mask estimation. The CNN is trained to learn the mapping between the noisy input signal's magnitude spectrogram and the Wiener filter, which is modeled to separate the direct path speech from directional non-stationary interference and reverberation. The room impulse responses (RIRs) from several rooms are used to generate speech signals for training a single CNN for the mask prediction in the presence of everyday interference sounds with varying levels. The CNN-based weighted SRP-PHAT is tested by localizing real moving and static speech sources in the presence of directional time-varying interference and reverberation. Results show that the directional likelihood of the speaker is increased over the traditional SRP-PHAT. Consequently, the number of correct speaker direction estimates is increased over the traditional SRP-PHAT.

The paper is organized as follows. Section 2 reviews the sig-

The authors wish to acknowledge CSC IT Center for Science, Finland, for computational resources.

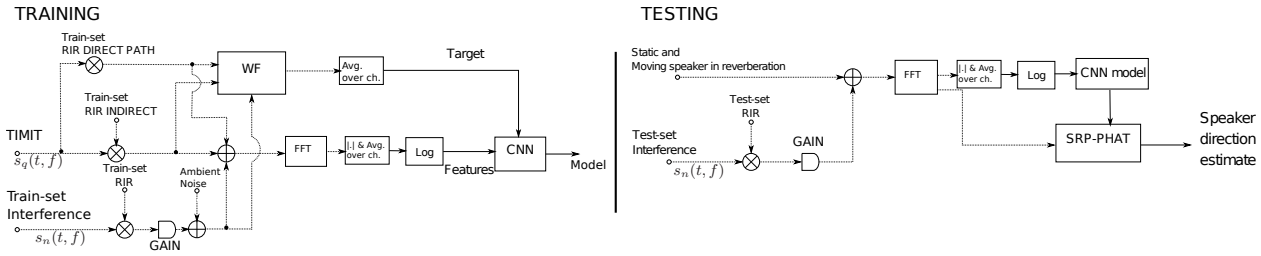

Fig. 1. Training and testing framework for the proposed system. WF represents Wiener filter and FFT represents Fast Fourier transform. Dashed lines indicate multi-channel input. "$|\cdot|$ & Avg. over ch." indicates averaging absolute values over channels.

nal model, presents the CNN-based TF mask learning system, and the training procedure. Section 3 describes the static simulated data, which is used to train the CNN-based TF mask learning, and the recorded speech sentences that are used to evaluate the localization performance. In Section 4 the sound source localization problem is reviewed using TF masking in the SRP-PHAT framework. Localization results are presented in Section 5, and Section 6 concludes the discussion.

2. CNN-BASED TF MASK LEARNING

This section presents the used signal model, and how to obtain the TF weights for the desired class of speech in presence of time-varying interference sources.

The i^{th} microphone signal $x_i(t,f)$ is expressed in the time-frequency domain, where $f = 0,\ldots,K-1$ is discrete frequency index and t is processing frame index. The signal is modeled as the sum of $n \geq 0$ reverberated signals $s_n(t,f)$ emitted from positions $\mathbf{r}_n$ in presence of noise $e_i(t,f)$

$$x_i(t,f) = \textstyle\sum_n h_{\mathbf{m}_i,\mathbf{r}_n}(f) \cdot s_n(t,f) + e_i(t,f), \qquad (1)$$

where $h_{\mathbf{m}_i,\mathbf{r}_n}(f)$ is the RIR between source position $\mathbf{r}_n \in \mathbb{R}^3$ and microphone position $\mathbf{m}_i \in \mathbb{R}^3$, both in Cartesian coordinates, and $i = 1,\ldots,M$, where M is the number of microphones.

The TF mask of i^{th} microphone signal is denoted as $\eta_i(t,f)$ and it takes values in range $[0,1]$. The mask is designed to reduce the contribution of TF points that do not belong to the target source [5]. Here, the Wiener filter is used to model the channel specific TF mask of desired source q

$$\eta_i(t,f) = \frac{|h^{\text{dp}}_{\mathbf{m}_i,\mathbf{r}_q}(f) \cdot s_q(t,f)|^2}{|h^{\text{dp}}_{\mathbf{m}_i,\mathbf{r}_q}(f) \cdot s_q(t,f)|^2 + |u_i(t,f)|^2}, \qquad (2)$$

where $|\cdot|$ is absolute value, $h^{\text{dp}}_{\mathbf{m}_i,\mathbf{r}_q}(f)$ is the direct path-only component of the RIR, and $u_i(t,f)$ is the mixture of all undesired signals, i.e., the reverberated interference sources and target signal reverberation

$$\begin{aligned} u_i(t,f) &= \bar{h}_{\mathbf{m}_i,\mathbf{r}_q}(f) \cdot s_q(t,f) \\ &\quad + \sum_{n \neq q} h_{\mathbf{m}_i,\mathbf{r}_n}(f) \cdot s_n(t,f) + e_i(t,f), \end{aligned} \qquad (3)$$

and $\bar{h}_{\mathbf{m}_i,\mathbf{r}_q}(f)$ denotes the RIR without the direct path component i.e. it models only the undesired reverberation of the target source. The latter sum of Eq. (3) is the sum of undesired interference sources convolved with their corresponding RIRs in presence of added noise.

The input to the CNN is a 32-frame patch of the log-magnitude spectrogram. The target output is the corresponding TF mask patch. To save computational cost of training channel specific masks, the training target is obtained by averaging the masks over the array channels. Similarly, the input feature is averaged over the channels. The resulting mask is therefore common for all channels, i.e. $\eta_i(t,f) = \eta_j(t,f),\ \forall i,j,t,f$.

The CNN architecture used in this work are as follows. Four convolutional layers with 96 feature maps and rectified linear unit (ReLU) activation functions perform 11-by-5 (time-by-frequency) convolution over their inputs. First three convolutional layers are followed by a max-pooling layer with downsampling pool size of four frequency bins. Max-pooling is not performed over time domain, as our aim is to obtain the estimated TF mask for each frame. The convolutional layers are followed by an output feed-forward layer with sigmoid activation. The input for this layer is the concatenated features from each feature map for each frame. Same feedforward layer weights are applied to the features from each frame. The sigmoid output of this layer is used as the estimated TF mask.

CNN training settings used in this work are as follows. Mean squared error is used as the loss function. For each convolutional layer, batch normalization [16] and dropout [17] with rate 0.25 is used. During training, Adam gradient-based optimization [18] is used. The optimal network parameters are found over grid search and the model is chosen as the one with the least average validation loss over different interference conditions.

3. DATA DESCRIPTION

The used RWCP-SSD database [19] consists of acoustically different spaces[1] with different reverberation characteristics. A room specific number of annotated source angles is used to capture static RIRs using a 16 microphone circular array with a 15 cm radius placed at approximately 2 m distance from the sources. In addition, a number of sentences spoken in Japanese are played back from a static and a moving loudspeaker. In the moving speaker recording, the speaker's azimuth angle follows a path between $50^\circ - 130^\circ$, during which the loudspeaker distance ranges from 1.6 m to 2.0 m from the microphone array. An infrared-based tracking device is used to produce ground truth direction of the moving speaker. The static speech recordings utilize a single loudspeaker angle, and the ground truth angle is obtained from the acoustic signal[2]. The amount of reverberation, number of used RIRs, and the number of available speech recordings is presented in Table 1 for each used room. The RIRs are divided into non-overlapping training, validation and testing sets. The data is mixed at 48 kHz sampling rate, which is the sample rate used to obtain RIRs. The generated mixtures are then downsampled to 16 kHz.

[1] Moving panels are used in order to alter reverberation to generate some of the spaces in addition to using different rooms.

[2] The median of azimuth angles obtained with Eq. (5) from frames with detected voice activity.

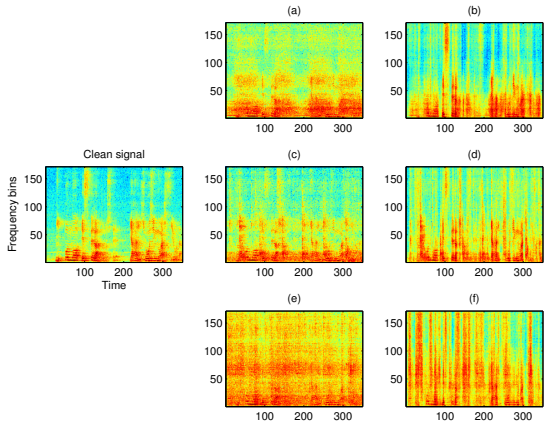

Fig. 2. Left: clean signal recorded from a moving speaker in reverberated conditions (no interference). Center and right: clean signal + 0 dB interference + CNN-based mask with interference source as (a,b): Household, (c,d): Interior background, (e,f): Printer noise.

3.1. Audio mixture generation for CNN training

The synthesized training data is produced as follows. Three sets of data is generated for CNN training, validation, and testing. Speech sentences from TIMIT database are divided based on speaker identity into these sets. Everyday sounds that consist of office printer noise, household noise, and interior background are used as interference, and are from the BBC sound effects library (obtained from *Stockmusic*[3]). Each class of interference signal is also divided into the three sets of non-overlapping recordings.

For each set, randomly selected speech signals are convolved with the set-specific RIRs corresponding to available loudspeaker angles. Interference signals are then convolved with the set-specific RIRs associated to another loudspeaker angle that is spatially non-overlapping and the signals are then mixed. The process is repeated for each room using all pairwise loudspeaker angle combinations. The speech-to-interference ratio (SIR), calculated using time-domain signal values, of -6 dB, 0 dB, $+6$ dB, and $+12$ dB are used to vary the level of interference. Finally, recorded ambience from room OFC is added with randomly drawn relative level between $[-6, -12]$ dB with respect to the observed speech level to model realistic room conditions. Five repetitions of the described process is used to produce a total 22440 training, 3000 validation, and 5160 test mixtures for the CNN training.

3.2. Audio for localization performance analysis

The used speech data consists of array recordings of Japanese sentences emitted using either the static or the moving loudspeaker. These speech recordings are mixed with the (test-set) interference signals using all of the test-set RIRs. These signals or the utilized RIRs have not been used to train the CNN. Prior to mixing, the speech recordings are pre-processed by spectral subtraction to decrease the amount of static noise using a quantile-based method [20] (with $q = 0.3$). The process is repeated five times using the same SIR levels as in CNN training data generation. A total of 1440 and 1560 mixtures are created to evaluate the localization performance of the moving and static speaker, respectively.

[3] www.stockmusic.com

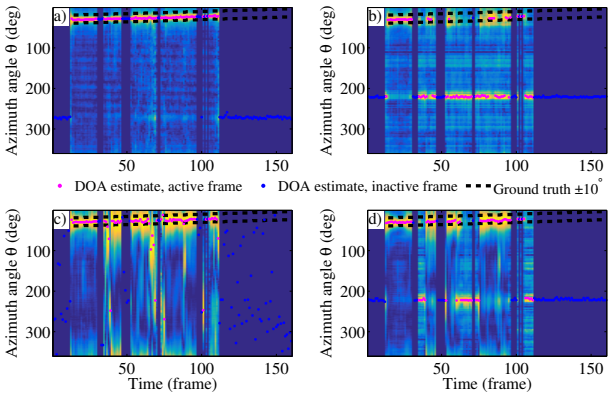

Fig. 3. a) SRP-PHAT $L(\theta, t)$ for moving speaker in reverberation. b) SRP-PHAT with added printer interference at +6 dB SIR, c) SRP-PHAT with ICM weight for the mixture, d) SRP-PHAT with CNN predicted weight for the mixture. Black dashed lines represent the $\pm 10^\circ$ angle around the ground truth. Dots represent DOA estimates with color associated to speech activity. SRP values in frames without speech activity are set to zero.

4. SRP-PHAT WITH TF MASKING

Based on the weighted GCC-PHAT [5], the weighted SRP-PHAT is the sum of weighted GCC-PHAT functions for sound wave direction $\mathbf{k}$ over all microphone pairs $\{i, j\}$

$$L(\mathbf{k},t) = \sum_{i,j}\sum_{f=0}^{K-1} \frac{\eta_i(t,f)x_i(t,f)\cdot(\eta_j(t,f)x_j(t,f))^*}{|x_i(t,f)|\cdot|x_j^*(t,f)|} e^{\jmath\cdot\tau_{i,j}\cdot\omega_f}, \tag{4}$$

where $\eta_i(t,f)$ is the TF mask for the i^{th} signal, $\jmath$ is the imaginary unit, $(\cdot)^*$ denotes complex conjugate, $\omega_f = 2\pi f/K$ is the angular frequency, and $\tau_{i,j} = \mathbf{k}^T(\mathbf{m}_i - \mathbf{m}_j)/c$ is propagation time difference between microphones, where $\mathbf{k} \in \mathbb{R}^3$ is defined as a Cartesian unit vector of the sound wave direction and c is the speed of sound. The point estimate for DOA is obtained by

$$\hat{\mathbf{k}}(t) = \arg\max_{\mathbf{k}} L(\mathbf{k}, t). \tag{5}$$

5. RESULTS

The array signal is processed at a sample-rate of 16 kHz. The frame length is set to 21.4 ms. Sine windowing with 50 % overlap is applied. Refer to Fig. 1 for an overview of the CNN training and testing process.

To illustrate the interference reduction capability of the CNN-based mask, Fig. 2 displays an example of the moving speaker signal's spectrogram in three types of used interference, before and after applying the predicted mask.

Table 1. Reverberation times, amount of RIRs used for training, validation and testing, and the number of speech recordings is specified for each of the rooms. Room naming conventions follow the RWCP-SSD database [19].

Room	E2A	E2B	E1B	E1C	OFC	JR1	JR2
RT_{60}(s)	0.3	1.3	0.31	0.38	0.78	0.6	0.47
Train RIRs	8	8	9	9	8	8	3
Validation RIRs	3	3	4	4	3	3	2
Test RIRs	3	3	6	6	3	3	2
Moving speakers	50	50	0	0	50	50	0
Static speakers	50	50	0	0	50	50	50

Table 2. Relative amount of SRP-PHAT likelihood mass near the ground truth source angle ($\pm 10^\circ$) normalized with SRP-PHAT of the speech signal without interference. Difference to SRP-PHAT is given for CNN and ICM-based weighting.

SIR[dB]	Household			Interior background			Print		
	SRP-PHAT weight			SRP-PHAT weight			SRP-PHAT weight		
	-	ΔICM	ΔCNN	-	ΔICM	ΔCNN	-	ΔICM	ΔCNN
Static speaker									
+12	64.1	+6.5	+10.1	69.8	+8.6	+12.8	54.8	+7.4	+11.7
+6	53.8	+8.4	+9.3	58.9	+10.7	+12.2	45.8	+9.9	+9.7
0	46.1	+10.4	+7.3	48.8	+13.4	+10.6	36.7	+12.1	+6.3
−6	36.2	+13.1	+4.2	38.9	+14.6	+7.0	29.8	+13.6	+3.4
Moving speaker									
+12	68.6	+5.5	+10.9	73.5	+7.6	+12.8	58.7	+6.7	+12.1
+6	59.0	+7.4	+9.7	63.0	+10.5	+13.0	48.5	+10.0	+9.8
0	50.2	+10.2	+7.3	52.1	+13.2	+10.5	39.9	+12.8	+6.5
−6	38.5	+13.7	+4.4	41.4	+15.4	+6.3	32.0	+15.2	+3.5

Speaker localization performance is analyzed by examining relative frame-wise SRP-PHAT likelihood around $\pm\psi$ azimuth angle of the ground truth source direction[4] $\theta(t)$ in one degree resolution

$$p_\psi(t) = \Big(\sum_{\vartheta=-\psi}^{\psi} L(\theta(t)+\vartheta,t)\Big) / \Big(\sum_{\vartheta=0^\circ}^{359^\circ} L(\vartheta,t)\Big). \quad (6)$$

The SRP-PHAT values are normalized in each frame separately by first removing the minimum value and then by dividing with the maximum value. Three SRP-PHAT variants are evaluated for the mixtures using source angles in the horizontal plane:

1. Traditional SRP-PHAT (i.e. $\eta_i(t,f)=1$), refer to Eq. (4)
2. Proposed CNN-based mask weighted SRP-PHAT, refer to Sections 2 and 4.
3. Interference canceling mask (ICM) weighted SRP-PHAT, refer to Eqs. (2), and (4). The ICM is obtained using the WF, where the reverberated interference signal is the noise signal, and the array recording of the speech is the target signal.

Figure 3 panel a) depicts the SRP-PHAT for the moving speaker without interference, panel b) illustrates SRP-PHAT output for the mixture signal (printer) in +6 dB SIR case. Panels c) and d) depict the ICM and the CNN mask weighted SRP-PHAT, respectively. The interference source is in direction 220° and its SRP-PHAT likelihood is decreased by the use of masks, refer to panels c) and d). Note also the restoration of correct DOA estimates as a result of applying the CNN mask, e.g., in frames 70–100.

Only frames with detected voice activity are included in the analysis. Frame level voice activity is estimated by a feed forward neural network that was trained using MFCC features from a sub-set of test-data, where target values were obtained with manual labeling.

Table 2 presents the relative amount of SRP-PHAT likelihood mass near the ground truth source direction ($\pm 10^\circ$ azimuth). The portion of SRP-PHAT likelihood near ground truth source direction without interference is used as the reference value of 100%. Difference to basic SRP-PHAT is given for the ICM and CNN-based methods. The results for each combination of interference signal type and SIR level are given as average values over different rooms, interference source angles, and repetitions.

For every type of interference in every SIR level, in contrast to SRP-PHAT, the proposed CNN masking results in higher concentration of SRP-PHAT likelihood around the ground truth source direction for both static and moving speakers. In every +12 dB and most of the +6 dB SIR mixtures the CNN masking even surpasses the ICM. Since here the ICM has only access to the reverberated captured speech and not the direct path signal it is plagued with the undesired reflections of speech from walls. These reflections contribute to the SRP-PHAT evidence in regions outside of source directions. In contrast, the CNN mask is trained to separate the direct path component from reflections and is able to suppress the low level of interference and the reverberation. However, in higher levels of interference the ICM results in better localization performance than the mask predicted by the CNN. This is most likely due to non-ideal generalization performance of the CNN to unseen data, but the results are still improved over traditional SRP-PHAT.

[4]For simplicity, we denote the SRP-PHAT function $L(\mathbf{k},t)$ of Eq. (4) with only azimuth angle $L(\theta,t)$ when the elevation angle is in the horizontal plane. The relationship between DOA vector $\mathbf{k}$ and azimuth angle θ is Cartesian to spherical transformation.

Table 3. DOA point estimate results relative to SRP-PHAT results without interference. Percentage of correct DOA estimates around ground truth source angle ($\pm 10^\circ$) in the presence of interference. Difference to SRP-PHAT is given for CNN and ICM-based methods.

SIR[dB]	Household			Interior background			Print		
	SRP-PHAT weight			SRP-PHAT weight			SRP-PHAT weight		
	-	ΔICM	ΔCNN	-	ΔICM	ΔCNN	-	ΔICM	ΔCNN
Static speaker									
+12	67.1	+11.1	+3.2	75.1	+8.3	+2.7	55.1	+12.8	+6.3
+6	51.3	+15.9	+5.5	60.7	+13.9	+4.1	38.8	+19.0	+7.9
0	36.6	+20.8	+6.8	44.5	+20.3	+6.4	19.3	+25.3	+9.2
−6	18.8	+25.0	+7.1	25.7	+24.6	+6.1	6.7	+25.9	+8.0
Moving speaker									
+12	73.8	+9.0	+2.9	80.2	+7.3	+2.3	63.5	+11.3	+5.2
+6	59.9	+14.5	+4.8	68.3	+13.1	+3.7	44.8	+19.8	+7.4
+0	44.8	+20.7	+5.4	51.7	+20.0	+4.9	25.9	+27.6	+9.5
−6	23.5	+28.3	+7.8	31.5	+27.1	+4.3	9.7	+31.1	+8.8

Table 3 presents the percentage of correct point DOA estimates for the three SRP-PHAT variants over frames with speech activity. The angle resulting in maximum SRP-PHAT likelihood is taken as the point estimate for each frame, refer to Eq. (5). The number of correct DOA estimates for SRP-PHAT in the ground truth direction without interference is used as the reference value of 100 %. Difference to basic SRP-PHAT is given for the ICM and CNN-based methods. The ICM has the best performance in all SIR levels and for all interference types. The proposed CNN masking retains a higher percentage of correct DOA values than the basic SRP-PHAT in all cases.

6. CONCLUSIONS

This paper considers using time-frequency masks, predicted by CNNs, for the reduction of the detrimental effects caused by reverberation and time-varying interference in the SRP-PHAT localization function. Measured RIRs, speech, and interference signals are used to construct training data for the CNN in order to learn the mapping between a noisy input log-magnitude spectrogram and a corresponding desired TF mask. The mask is here modeled using the Wiener filter to reduce the magnitude of the interference and the reverberation of the target speech signal.

The proposed method is tested using both static and moving speech sources mixed with different levels of everyday interference signals. Averaged over different rooms, interference source angles, and repetitions, the CNN-based TF mask was found to reduce the effects of interference and reverberation in the SRP-PHAT likelihood map. This is also evident in the increased number of DOA estimates towards the ground truth source direction. The localization results were obtained using Japanese speech, while training was performed with English (TIMIT sentences). This mismatch suggests a good generalization capability of the CNN-based mask estimator.

7. REFERENCES

[1] Xinya Li, Zhiqun Daniel Deng, Lynn T. Rauchenstein, and Thomas J. Carlson, "Contributed review: Source-localization algorithms and applications using time of arrival and time difference of arrival measurements," *Review of Scientific Instruments*, vol. 87, no. 4, 2016.

[2] Benoit Champagne, Stéphane Bédard, and Alex Stéphenne, "Performance of time-delay estimation in the presence of room reverberation," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 2, pp. 148–152, 1996.

[3] Kevin W. Wilson and Trevor Darrell, "Learning a precedence effect-like weighting function for the generalized cross-correlation framework," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2156–2164, Nov 2006.

[4] Jean-Marc Valin, Francois Michaud, and Jean Rouat, "robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering," *"Robotics and Autonomous Systems"*, vol. 55, no. 3, pp. 216–228, 2006.

[5] François Grondin and François Michaud, "Time difference of arrival estimation based on binary frequency mask for sound source localization on mobile robots," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 6149–6154.

[6] Cha Zhang, Dinei Florêncio, and Zhengyou Zhang, "Why does phat work well in low noise, reverberative environments?," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 2565–2568.

[7] DeLiang Wang and Guy J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press, 2006.

[8] Arun Narayanan and DeLiang Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7092–7096.

[9] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[10] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] Tom Sercu, Christian Puhrsch, Brian Kingsbury, and Yann LeCun, "Very deep multilingual convolutional neural networks for lvcsr," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4955–4959.

[12] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *arXiv preprint arXiv:1604.07160*, 2016.

[13] Viren Jain and Sebastian Seung, "Natural image denoising with convolutional networks," in *Advances in Neural Information Processing Systems*, 2009, pp. 769–776.

[14] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[15] Like Hui, Meng Cai, Cong Guo, Liang He, Wei-Qiang Zhang, and Jia Liu, "Convolutional maxout neural networks for speech separation," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2015, pp. 24–27.

[16] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[17] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] Satoshi Nakamura, Kazuo Hiyane, Futoshi Asano, Takanobu Nishiura, and Takeshi Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *2nd International Conference on Language Resources & Evaluation, (LREC)*, 2000.

[20] Volker Stahl, Alexander Fischer, and Rolf Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2000, vol. 3, pp. 1875–1878.