Deep Neural Networks for joint Voice Activity Detection and Speaker Localization

Paolo Vecchiotti, Emanuele Principi, Stefano Squartini, Francesco Piazza
Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy

***Abstract*—Detecting the presence of speakers and suitably localize them in indoor environments undoubtedly represent two important tasks in the speech processing community. Several algorithms have been proposed for Voice Activity Detection (VAD) and Speaker LOCalization (SLOC) so far, while their accomplishment by means of a joint integrated model has not received much attention. In particular, no studies focused on cooperative exploitation of VAD and SLOC information by means of machine learning have been conducted, up to the authors' knowledge. That is why the authors propose in this work a data driven approach for joint speech detection and speaker localization, relying on Convolutional Neural Network (CNN) which simultaneously process LogMel and GCC-PHAT Patterns features. The proposed algorithm is compared with a two-stage model composed by the cascade of a neural network (NN) based VAD and an NN based SLOC, discussed in previous authors' contributions. Computer simulations, accomplished against the DIRHA dataset addressing a multi-room acoustic environment, show that the proposed method allows to achieve a remarkable relative reduction of speech activity detection error equal to 33% compared to the original NN based VAD. Moreover, the overall localization accuracy is improved as well, by employing the joint model as speech detector and the standard neural SLOC system in cascade.**

I. Introduction

Nowadays, in many engineering systems commonly integrated in our life, the extraction and processing of contextual information coming from multimedia signals is widely used. In many applications, the speech utterances emitted by people in the environment under study often constitute one of the most representative information sources in this sense, and diverse *speech processing* algorithms can be employed on purpose, in dependence on the specific tasks under investigation.

The focus in this work is on Voice Activity Detection (VAD) and Speaker LOCalization (SLOC). The first one plays a fundamental role in several scenarios, such as mobile phone communication, echo cancellation, and speech recognition. Similarly, localization algorithms deserve attention in the development of different tasks, as monaural or binaural models based on the human hearing system [1], or the interaction between human and robots [2].

Different techniques have been proposed in the literature to tackle the voice activity detection problem in indoor environments. Among the most recent ones, an approach recognizing a reference *anchor* word with the help of mean subtraction is discussed in [3], the interaction between VADs based on the SNR estimate is investigated in [4]. Deep neural networks have been employed in [5], in which the authors proposed a neural network based VAD, focusing on multi-stage optimization. Similarly, Convolutional Neural Networks (CNN) with 3-D kernels have been used in [6]. At the same time, several approaches have been proposed for localizing a speaker in closed environments. In [7], the localization algorithm takes advantage of the signal energy measure, Direction Of Arrival of the audio signal is estimated in [8], while [9] employs the Steered-Response Power Phase Transform (SRP-PHAT). In addition, the SLOC problem has been recently faced by means of neural networks in [10], [11], especially with a focus on CNN [2], [12]–[14].

In the last years, the simultaneous detection and localization of a speaker has been addressed in different works. Commonly, VAD and SLOC are disposed as a cascade [15]–[18] or a parallel [19] configuration. Up to the authors' knowledge, only two contributions investigate the cooperation between VAD and SLOC. One is the approach proposed in [17], in which an ensemble integration of speaker localization and statistical speech detection data in domestic environments is implemented. The second technique jointly performs VAD and SLOC [20] by employing a modified version of SRP-PHAT algorithm.

However, a single data-driven model for joint speaker detection and localization has never been investigated. Therefore, this work is intended to simultaneously exploit both VAD and SLOC data in order to improve the overall performance, both in terms of speech detection and speaker localization. Deep neural networks (DNN) are employed on purpose, for two main reasons. First, DNN have already shown remarkable performance on the two separate tasks, as mentioned above. Second, a neural architecture with its multiple inputs and outputs allows to easily make use of VAD and SLOC feature data and decision variable values.

In this work, a classic cascade configuration is firstly developed, where a neural SLOC is trained by means of an Oracle VAD selecting only the speech portions of audio signals. A real VAD is employed for evaluating the SLOC performance. In details, speaker localization error is computed on true positive speech frames detected by the VAD. Subsequently, a new model based on CNN, simultaneously operating as detector and localizer exploiting standard VAD and SLOC features, is proposed. The training of this network is performed by using both speech and non-speech signals.

For the proposed study, the multi-room scenario already

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU used for this research.

ISBN 978-90-827970-1-5 © EURASIP 2018

addressed in the authors' previous contribution [14] is taken as reference, in order to have a solid experimental background for evaluating the proposed approach. Accomplished computer simulations show that the joint model allows to remarkably improve the VAD performance, with respect to the original non-cooperative solution. On the other side, the localization ability is enhanced when the standard neural network based SLOC is piloted with the proposed joint model operating as VAD.

The paper is organized as follows. In Section II the proposed algorithm is described. The computer simulation setup is presented in Section III, whereas the experimental results are reported in Section IV. Section V concludes the work.

II. Proposed Method

In this section, the comparative model is presented in the first place, being the cascade of the so-called *Neural VAD* and *Neural SLOC*, already proposed by the authors in their previous works [6] [14]. Then, the description of the new proposed method named *Joint VAD-SLOC Model* is given. Finally, the shared details of two models are briefly illustrated.

A. Cascade Model

This model is made by the cascade of the Neural VAD and Neural SLOC. Speech detection is performed by the former. It consists in a CNN fed by LogMel features extracted from all the available microphones. Training and testing of Neural VAD is accomplished over speech and non speech data acquired by means of environmental microphones. The Neural SLOC is formed by a CNN processing GCC-PHAT Patterns. Localizing a speaker is dealt with as a 2-D problem, where the height of the speaker from the floor is not considered. The 2-D speaker coordinates defined as $(\chi\ ,\ \psi)$ are directly predicted by the CNN. An Oracle VAD selecting only speech frames is used during the training phase of the Neural SLOC, as in [14]. In computer simulations, as discussed later on, the Neural SLOC has been tested using only speech frames detected by Oracle VAD and by Neural VAD, i.e., considering all the available speech frames in the dataset and the true positive predictions of the Neural VAD, respectively.

B. Joint VAD-SLOC Model

In this neural model, the simultaneous detection of speech frames and localization of speaker position is performed. As discussed in the introductory section, the objective is to exploit the synergy between these two tasks to improve their performance, and a full data-driven technique was identified as the most viable solution to implement the idea. Several options have been investigated, and the most performing one is the model depicted in Fig. 1. It consists in a single CNN with two separate stacks of convolutional layers separately processing LogMel and GCC-PHAT Patterns features, followed by a common set of standard feed-forward layers. The network ends with three outputs, being the voice activity prediction and the two speaker position coordinates. In details, as it will be described in Section II-D, also the two coordinates are eligible to represent the speech/ non-speech condition. A specific threshold needs to be used on purpose, as discussed in Section II-E. This last strategy has been adopted, indeed. Moreover, since both speech and non-speech frames are used to train the model, the authors propose to label the speaker position in the non speech condition as a physical location outside the considered room, as similarly done in [20].

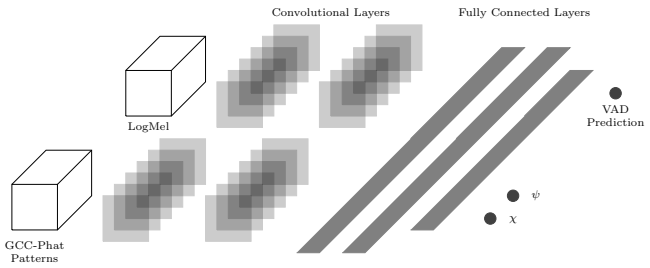

Figure 1 The Convolutional Neural Network employed for the Joint VAD-SLOC Model. Pooling layers are absent. The outputs of the network are three neurons, one for speech detection and the other two for speaker localization.

C. Features

1) GCC-PHAT Patterns: This kind of feature aims to estimate the time difference occurring between the two signals captured by a microphone pair in presence of a sound event. GCC-PHAT Patterns features is the result of the frequency domain cross correlation of these two signals. Only adjacent microphones pairs are chosen for features extraction. Plus, due to the spatial disposition of the microphones, the first 51 values of the cross correlation are selected. Signals are sampled at 16 kHz, while frame size and hop size are set to 30 ms and 10 ms respectively. Mean and variance normalization is applied. The authors have already employed GCC-PHAT Patterns in [14] for the SLOC task.

2) LogMel: This feature set is widely employed in computational audio processing. The authors used it in [6] for the VAD task. The LogMel features are calculated as follows: the signal frame goes through the Fourier transform, 40 mel-band filters are applied and, finally, the logarithm of the power spectrum is calculated for each mel-band. Frame size is chosen as 25 ms and hop size 10 ms.

3) Temporal Context: An improvement of CNN performance has been observed in authors' previous work [14] by extending the processed input data including also past and future occurrences. The same approach has been used here as well. Two are the parameters to set in this case, i.e., *context* and *strides*. The first indicates the total number of frames considered as input instead of the single actual frame, where an equal number of past and future frames is selected. The latter pilots the selection of previous and future frames.

D. Convolutional Neural Network

CNN has encountered a large success in computational audio processing applications in the recent years. The authors used this kind of network in all neural models addressed in this work, i.e., the Neural VAD, the Neural SLOC and the Joint VAD-SLOC Model. Typically, a 2-D convolution over the input matrix is performed by means of the CNN convolutional layers, generating a set of *feature maps*, then processed by

ISBN 978-90-827970-1-5 © EURASIP 2018

a stack of neuronal dense layers. As mentioned above, in the proposed Joint VAD-SLOC Model, the final outputs of the convolutional layers processing LogMel and GCC-PHAT Patterns are concatenated and then processed by the following layers of neurons.

1) Speech and Non-Speech Labelling: In the case of Neural VAD, a 0 or 1 label is used for speech/non-speech classification. For the Neural SLOC, following the authors' previous work [14], the neural network outputs represent the coordinates (χ, ψ) of the speaker inside the room, which range in $[0, 1]$, when speech frames are processed. The three outputs of the Joint VAD-SLOC Model are the combination of Neural VAD and Neural SLOC outputs. Nonetheless, while the label for VAD prediction is kept boolean as Neural VAD, the non speech frames lack of a label for the coordinates outputs. The two (χ, ψ) coordinates are labeled as -1 for non-speech frames.

2) Activation Function: For the Neural VAD and Neural SLOC case studies, the *ReLU* activation is employed. The need of a specific activation function for the Joint VAD-SLOC Model rises, due to localization labels ranging in $[-1, 1]$. The *Hard Tanh* nonlinearity has been chosen, which acts as $f(x) = x$ in $[-1, 1]$ and saturates to -1 and 1 out of this range.

E. Post Processing

1) VAD: A threshold is applied for discriminating speech and non-speech. In the case of the Neural VAD, its application is straightforward. Differently, when the localization coordinates are considered for VAD decision as described in Section II-B, the application of threshold becomes a 2-D problem. The authors propose to use a straight line in the position coordinate space as threshold, as depicted in Fig. 2. Subsequently, a standard *Hangover* technique is applied to the VAD prediction values. It employs a counter K: the actual frame is non speech if and only if the previous $(K-1)$ frames are non speech. K is set equal to 8.

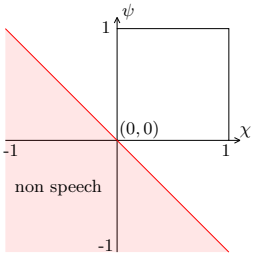

Figure 2 The application of the 2-D threshold. The square box top view of the room, where the walls are normalized in the range [0,1]. Speech is expected to be predicted in that box. The thin red line is the threshold. If prediction lies in the red region then it is labeled as non speech, otherwise it is considered as speech.

2) SLOC: The smoothing of the predicted coordinates is carried out by means of a moving average filter. The window size of the filter is set equal to 5.

III. Experimental Setup

A. DIRHA Dataset

The DIRHA dataset has been used in this study [21]. It consists of multiple recordings acquired by means of 40 microphones installed in the walls and the ceilings of a five rooms apartment, as depicted in Fig. 3. A distance of 50 cm occurs between two close microphones. The *Real* and the *Simulated* subsets compose the DIRHA dataset; the proposed approach is tested against the latter, which consists of 80 scenes lasting a minute each, with a total amount of speech equal to 23.6 minutes. More details are provided in [21]. Simulations address two of the five rooms, which are the Living Room and the Kitchen. These rooms are chosen since most of the speech events occurs there, plus a higher number of microphones is available.

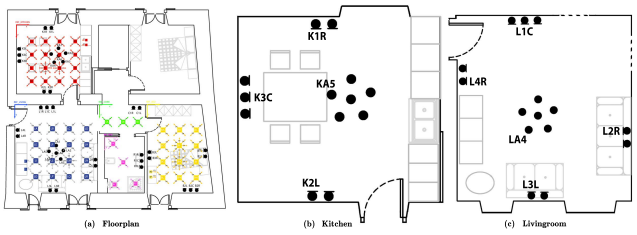

Figure 3 The map of the apartment used for the DIRHA project (a). Figures (b) and (c) show the considered rooms, where the thick black dots are the installed microphones.

1) Microphones Selection: A fixed set of microphones is employed in this study. For the Neural VAD, all the available microphones are considered, from which LogMel features are evaluated. Regarding the Neural SLOC, GCC-PHAT Patterns are extracted from all the couples of adjacent microphones installed in the wall and the ceiling array (i.e., microphones pairs distancing 50 cm). The central microphones of the ceiling arrays (KA6, LA6) are excluded. As mentioned above, the Joint VAD-SLOC Model relies on LogMel and GCC-PHAT Patterns. The first are extracted from a reduced set of microphones which are K1R, K2L, K3C, KA5 for the kitchen and L1C, L2R, L3L, L4R, LA5 for the living room. For the latter, the same strategy adopted for the Neural SLOC is followed.

B. Evaluation Metrics

Two main groups of metrics are employed, one for detection and the other for localization. VAD evaluation relies on the false alarm rate (FA), the deletion rate (Del) and the overall speech activity detection (SAD) defined as follows:

$$\text{Del} = \frac{N_{del}}{N_{sp}}, \quad \text{FA} = \frac{N_{fa}}{N_{nsp}}, \quad \text{SAD} = \frac{N_{fa} + \beta N_{del}}{N_{nsp} + \beta N_{sp}}, \quad (1)$$

where N_{del}, N_{fa}, N_{sp} and N_{nsp} are the total number of deletions (false negative), false alarms (false positive), speech and non-speech frames, respectively. The term $\beta = N_{nsp}/N_{sp}$ plays the role of regulator due to class unbalancing.

Localization accuracy is given in terms of Root Mean Square Error (RMSE) and P_{cor}. RMSE is defined as:

$$\text{RMSE} = \frac{\sum_{i=0}^{N_{TOT}} \sqrt{(\chi_i - \chi_{\text{ref},i})^2 + (\psi_i - \psi_{\text{ref},i})^2}}{N_{TOT}}, \quad (2)$$

where χ_i and ψ_i are the i-th network outputs, $\chi_{\text{ref,i}}$ and $\psi_{\text{ref},i}$ are the i-th reference speaker coordinates, and N_{TOT} is the total number of frames. The second metric is defined as $P_{cor} = N_{FINE}/N_{TOT}$, where N_{FINE} is the number of

ISBN 978-90-827970-1-5 © EURASIP 2018

frames localized with RMS inferior than 500 mm. Localization performance is assessed only in correspondence of speech frames.

C. Neural Networks details

1) Implementation and Training: All algorithms have been implemented in the Python language using Keras [22] as deep learning library. All the experiments were performed on a computer equipped with a 6-core Intel i7, 32 GB of RAM and two Nvidia GTX 970, 4 GB of RAM graphic cards. The CNN training is performed by using standard backpropagation with the *Adam* optimizer [23]; plus, early stopping and variable learning rate are employed. Details are reported in Table I.

2) Cross Validation: A 10-fold cross validation is employed, with 8 folds used of training, one for validation and one for testing. Thus, the 80 scenes of the DIRHA Simulated dataset are grouped in 10 subsets of 8 scenes each. The scene selection procedure here employed aims to balance the amount of speech between the 10 subsets. In particular, the scene with the maximum amount of speech is selected and allocated into the first subset, then discarded. The next scene is selected in the same way, and allocated into the second subset, and so forth. The speech balancing operated by this data folds organization has shown to improve the training convergence behavior of the neural models.

The CNN hyper-parameters optimization is executed by random search; a total of more than 30 neural architectures is investigated for each model. Context and strides have been chosen a priori, as follows: context is set to 15 in all cases, while strides is equal to 4 for Neural VAD, 5 for Neural SLOC and 3 for the Joint VAD-SLOC Model.

	Training Epochs	Early Stopping	Learning Rate
Neural VAD	30	10	1e−5
Neural SLOC	500	50	2.5e−4
Joint VAD-SLOC Model	500	50	2.5e−4

Table I CNN Training Parameters

IV. Results and Discussion

The CNNs employed in the two rooms for the Neural VAD are similar, counting two layers of 128 kernels sized 3×3, succeeded by two layers of neurons, being 1024, 1024 for the kitchen and 1024, 256 for the living room. ReLU is employed as activation. All results discussed in this section are obtained by choosing the best threshold in the different addressed case studies. In Table II the results of the Neural VAD are reported.

	Kitchen	Living Room	Average
SAD (%)	5.6	4.8	5.2
DEL (%)	6.8	5.7	6.2
FA (%)	4.5	4.0	4.2

Table II Results of the Neural VAD applied on the two considered rooms of the dataset.

The Neural SLOC employs a CNN consisting of 128 kernels sized 7×7 for the kitchen and 5×5 for the living room. For each room, the convolutional layers are followed by two layers composed of 1024, 256 units. The Neural SLOC is tested on speech detected by an Oracle VAD or by Neural VAD. When the latter is employed, the Neural SLOC accuracy increases, since it is tested against a reduced set of speech (true positive), instead of all the available speech. This means that the Neural VAD fails in detecting frames in which the Neural SLOC is less accurate.

		Kitchen	Living Room	Average
Oracle VAD	RMS (mm)	332	359	345
	P_{cor} (%)	76	77	76
Neural VAD	RMS (mm)	317	337	327
	P_{cor} (%)	77	78	77

Table III Performance of the Neural SLOC. Its test takes place over all the speech of the dataset detected by the Oracle VAD, or against the speech detected by the Neural VAD reported in Table II, i.e. only for true positive.

The Joint VAD-SLOC Model employs the same CNN topology for the two rooms. Two separated stacks of two convolutional layers process LogMel and GCC-PHAT Patterns, respectively. Each one of the four layer is composed by 64 kernels of size 5×5. Three fully-connected layers respectively with 1024, 1024, 256 units and Hard Tanh activation function follow the convolutional layers. The two coordinates are used for speech detection by using the threshold described in Section II-E; the VAD prediction is rejected being less accurate. In Table IV the performance of the Joint VAD-SLOC Model are shown for detection and localization. The proposed method acts as a remarkable detector, outperforming the Neural VAD. On the other hand, the joint speech and non-speech training results more challenging in terms of localization, leading to a less accurate localizer compared to the Neural SLOC trained with the sole speech frames coming from the Oracle VAD.

	Kitchen	Living Room	Average
SAD (%)	3.8	3.1	3.5
DEL (%)	4.5	3.9	4.2
FA (%)	3.1	2.4	2.8
RMS (mm)	601	657	629
P_{cor} (%)	64	68	66

Table IV Results for the Joint VAD-SLOC Model.

Finally, a comparison for the average results of the three models is reported in Table V. The most performing configuration is obtained using the Joint VAD-SLOC Model as speech detector with the Neural SLOC in cascade. In terms of detection, comparing the Neural VAD and the Joint VAD-SLOC Model, SAD is decreased from 5.2% to 3.5% when the latter is employed, corresponding to a relative reduction equal to 33%. In addition, a lower SAD means as well that a higher number of true positive ($+3.1\%$) is detected by the Joint VAD-SLOC Model. Then, when assessing the localization accuracy of the Neural SLOC on speech frames detected by the Joint VAD-SLOC Model (Table Vb), a P_{cor} relative improvement of $+1.3\%$ is observed against the Neural SLOC tested on speech frames detected by Neural VAD. The average RMS reduces from 329 mm to 318 mm, i.e., a relative reduction of 3.34%.

Nevertheless, in Table III it was previously observed that the accuracy of the Neural SLOC increases when less true positive are detected, i.e., the Neural VAD is employed instead of the Oracle VAD. Hence, when detection is performed by the Joint VAD-SLOC Model rather than the Neural VAD, it

ISBN 978-90-827970-1-5 © EURASIP 2018

is reasonable to expect a decay of localization performance. Interestingly, the opposite takes place. This result shows that the Neural VAD fails to detect a subset of speech which is straightforward to localize for the Neural SLOC. Conversely, the Joint VAD-SLOC Model detects those speech frames, thus proving that the proposed model is able to cooperatively exploit detection and localization data.

Detection	Neural VAD	Joint VAD-SLOC Model
SAD (%)	5.2	3.5
DEL (%)	6.2	4.2
FA (%)	4.2	2.8

(a)

Localization	Neural SLOC*	Joint VAD-SLOC Model	Neural SLOC†
RMS (mm)	327	629	318
P_{cor} (%)	77	66	78

(b)

Table V Comparison of the two proposed models. The shown results are averaged between the two considered rooms. In (a) the comparison in terms of detection. (b) shows localization performances. Neural SLOC* means the localizer operating on the speech frames detected by Neural VAD, whereas Neural SLOC† operates on the speech frames detected by Joint VAD-SLOC Model.

V. Conclusion

The joint speech detection and speaker localization problem is addressed in this work. The authors aim to cooperatively exploit VAD and SLOC data by means of a data-driven approach, in order to improve the overall performance of the system. The proposed model, namely Joint VAD-SLOC Model, consists in a 3 outputs CNN processing LogMel and GCC-PHAT Patterns features. The model training makes use of non-speech frames, which requires the inclusion of a new label representing the localization of absent speakers. Computer simulations have been performed by considering a multi-room acoustic scenario and the DIRHA dataset has been used on purpose. In terms of speech detection, the Joint VAD-SLOC Model is compared with the original Neural VAD system, already proposed by the authors, leading to a relative reduction of average SAD error equal to 33%. The cascade of the Joint VAD-SLOC Model used as VAD with the Neural SLOC has been evaluated in terms of localization performance. The cascade configuration leads to a 2.7% relative improvement in terms of RMSE compared to the Neural SLOC. The obtained results thus confirm the effectiveness of the proposed idea.

Future works will be targeted to the improvement of localization accuracy of the Joint VAD-SLOC Model, by employing new specific features and augmenting the available speech frames in the original dataset. Moreover, the generalization of the proposed model when applied to diverse acoustic environments will be investigated.

References

[1] C. Pang, H. Liu, J. Zhang, and X. Li, "Binaural sound localization based on reverberation weighting and generalized parametric mapping," *IEEE Trans. Audio, Speech, Language Process.*, vol. 25, no. 8, pp. 1618–1632, Aug 2017.

[2] W. He, P. Motlicek, and J.-M. Odobez, "Deep neural networks for multiple speaker detection and localization," *arXiv preprint arXiv:1711.11565*, 2017.

[3] R. Maas, S. H. K. Parthasarathi, B. King, R. Huang, and B. Hoffmeister, "Anchored speech detection." in *INTERSPEECH*, 2016, pp. 2963–2967.

[4] P. Giannoulis, A. Brutti, M. Matassoni, A. Abad, A. Katsamanis, M. Matos, G. Potamianos, and P. Maragos, "Multi-room speech activity detection using a distributed microphone network in domestic environments," in *Proc. of EUSIPCO*, 2015, pp. 1271–1275.

[5] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep neural networks for multi-room voice activity detection: Advancements and comparative evaluation," in *Proc. of IJCNN*, Vancouver, Canada, 24-29 Jul. 2016, pp. 3391–3398.

[6] P. Vecchiotti, F. Vesperini, E. Principi, S. Squartini, and F. Piazza, "Convolutional neural networks with 3-d kernels for voice activity detection in a multiroom environment," in *Multidisciplinary Approaches to Neural Computing*. Springer, 2018, pp. 161–170.

[7] M. Cobos, F. Antonacci, A. Alexandridis, A. Mouchtaris, and B. Lee, "A survey of sound source localization methods in wireless acoustic sensor networks," *Wirel. Commun. Mob. Com.*, vol. 2017, p. 24, 2017.

[8] A. Tsiami, A. Katsamanis, P. Maragos, and G. Potamianos, "Experiments in acoustic source localization using sparse arrays in adverse indoors environments," in *Proc of EUSIPCO*, Lisbona, Portugal, Sep 1-5 2014, pp. 2390–2394.

[9] H. Do, H. F. Silverman, and Y. Yu, "A real-time SRP-PHAT source location implementation using stochastic region contraction (SRC) on a large-aperture microphone array," in *Proc. of ICASSP*, vol. 1, 2007, pp. I–121.

[10] N. Ma, T. May, and G. J. Brown, "Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 12, pp. 2444–2453, 2017.

[11] D. Goodman and R. Brette, "Learning to localise sounds with spiking neural networks," in *Adv. Neural Inf. Process. Syst.*, 2010, pp. 784–792.

[12] E. L. Ferguson, S. B. Williams, and C. T. Jin, "Sound source localization in a multipath environment using convolutional neural networks," *arXiv preprint arXiv:1710.10948*, 2017.

[13] S. Chakrabarty and E. A. Habets, "Multi-speaker localization using convolutional neural network trained with noise," *arXiv preprint arXiv:1712.04276*, 2017.

[14] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Localizing speakers in multiple rooms by using deep neural networks," *Computer Speech & Language*, vol. 49, pp. 83–106, 2018.

[15] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *Proc. of AVSS*, 2007, pp. 21–26.

[16] T. May, S. van de 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/ACM Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 2016–2030, 2012.

[17] Y. Tachioka, T. Narita, S. Watanabe, and J. Le Roux, "Ensemble integration of calibrated speaker localization and statistical speech detection in domestic environments," in *Proc. of HSCMA, 2014*, Florence, Italy, May 12-14 2014, pp. 162–166.

[18] R. Chakraborty and C. Nadeu, "Joint model-based recognition and localization of overlapped acoustic events using a set of distributed small microphone arrays," *arXiv preprint arXiv:1712.07065*, 2017.

[19] K. Lopatka, J. Kotus, and A. Czyzewski, "Detection, classification and localization of acoustic events in the presence of background noise for acoustic surveillance of hazardous situations," *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 10 407–10 439, 2016.

[20] M. J. Taghizadeh, P. N. Garner, H. Bourlard, H. R. Abutalebi, and A. Asaei, "An integrated framework for multi-channel multi-source localization and voice activity detection," in *Proc. of HSCMA*, 2011, pp. 92–97.

[21] L. Cristoforetti, M. Ravanelli, M. Omologo, A. Sosi, A. Abad, M. Hagmüller, and P. Maragos, "The DIRHA simulated corpus," in *Proc. of LREC*, vol. 5, Reykjavik, Iceland, May 26-31 2014.

[22] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Proc. of ICLR*, 2014.

ISBN 978-90-827970-1-5 © EURASIP 2018