
Article:
Vecchiotti, P., Pepe, G., Principi, E. et al. (2019) Detection of activity and position of speakers by using deep neural networks and acoustic data augmentation. Expert Systems with Applications, 134. pp. 53-65. ISSN: 0957-4174

https://doi.org/10.1016/j.eswa.2019.05.017

Detection of activity and position of speakers in real-world environments by using Deep Neural Networks and Acoustic Data Augmentation

Paolo Vecchiotti[a,*], Giovanni Pepe[a], Emanuele Principi[a], Stefano Squartini[a]

[a]*Department of Information Engineering, Università Politecnica delle Marche, via Brecce Bianche 12, 60131, Ancona, Italy*

Abstract

The task of Speaker LOCalization (SLOC) has been the focus of numerous works in the research field, where SLOC is performed on pure speech data, requiring the presence of an Oracle Voice Activity Detection (VAD) algorithm. Nevertheless, this perfect working condition is not satisfied in a real world scenario, where employed VADs do commit errors. This work addresses this issue with an extensive analysis focusing on the relationship between several data-driven VAD and SLOC models, finally proposing a reliable framework for VAD and SLOC. The effectiveness of the approach here discussed is assessed against a multi-room scenario, which is close to a real world environment. Furthermore, up to the authors' best knowledge, only one contribution proposes a unique framework for VAD and SLOC acting in this addressed scenario; however this solution does not rely on data-driven approaches.

This work comes as an extension of the authors' previous research addressing the VAD and SLOC tasks, by proposing numerous advancements to the original neural network architectures. In details, four different models based on convolutional neural networks (CNNs) are here tested, in order to easily highlight the advantages of the introduced novelties. In addition, two different CNN models go under study for SLOC. Furthermore, training of data-driven models is here improved through a specific data augmentation technique. During this procedure, the room impulse responses (RIRs) of two virtual rooms are generated from the knowledge of the room size, reverberation time and microphones and sources placement. Finally, the only other framework for simultaneous detection and localization in a multi-room scenario is here taken into account to fairly compare the proposed method.

As result, the proposed method shows to be more accurate than the baseline framework, and remarkable improvements are specially observed when the data

*Corresponding author

Email addresses: p.vecchiotti@pm.univpm.it (Paolo Vecchiotti), gpepe@pm.univpm.it (Giovanni Pepe), e.principi@univpm.it (Emanuele Principi), s.squartini@univpm.it (Stefano Squartini)

augmentation techniques is applied for both the VAD and SLOC tasks.

Keywords: voice activity detection, speaker localization, data augmentation, multi-room environment, deep learning

1. Introduction

The tasks of detecting human speech and the speaker position are respectively referred as Voice Activity Detection (VAD) and Speaker LOCalization (SLOC). In the research community, both deserve much attention, finding applications in audio surveillance, human hearing modelling, speech enhancement, human and robot interaction and so forth (Hughes & Mierle, 2013; Silva, Stuchi, Violato & Cuozzo, 2017; Tachioka, Narita, Watanabe & Le Roux, 2014; Taghizadeh, Garner, Bourlard, Abutalebi & Asaei, 2011; Vecchiotti, Principi, Squartini & Piazza, 2018). In the literature, speaker detection and its localization are generally treated as two separated problems. This strategy has led to numerous and efficient solutions for these two tasks. Indeed, in the early stages, a set of so-called classical VAD and SLOC algorithms was investigated. In particular, classical VADs are piloted by the analysis of specific signal characteristics (Benyassine, Shlomot, Su, Massaloux, Lamblin & Petit, 1997; Yantorno, Krishnamachari, Lovekin, Benincasa & Wenndt, 2001) or rely on statistical models of the speech and noise signals (Sohn, Kim & Sung, 1999; Lee, Nakamura, Nisimura, Saruwatari & Shikano, 2004). Similarly, the more general sound localization task has been tackled by classical techniques such as Cross Spectrum Phase (CSP) (Knapp & Carter, 1976) and Steered-Response Power Phase Transform (SRP-PHAT) (Do, Silverman & Yu, 2007; Seewald, Gonzaga Jr, Veronez, Minotto & Jung, 2014; Belloch, Gonzalez, Vidal & Cobos, 2015). These techniques rely on two main stages: initially cross-correlation is employed for estimating the Time Difference of Arrival (TDOA) between each microphone pair under study, then TDOAs are combined and jointly processed for localizing the sound source.

Recently, the study of VAD and SLOC algorithms has been heavily influenced by the break-through of deep learning and data-driven approaches. Indeed, the effectiveness of artificial deep neural networks (DNN) has been shown in different acoustic scenarios.

With regards to VAD, numerous DNN architectures have been investigated in the last years. A Recurrent Neural Network (RNN) model for VAD outperforms a Gaussian Mixture Model (GMM) in (Hughes & Mierle, 2013). For the multi-room domestic scenario numerous DNN-based VAD are discussed in (Ferroni, Bonfigli, Principi, Squartini & Piazza, 2015), where a Deep Belief Network achieves the highest accuracy compared to a Multi Layer Perceptron (MLP) and a Bidirectional Long Short-Term Memory (BLSTM) recurrent neural network. Furthermore, convolutional neural networks (CNNs) directly process the audio spectrogram in (Silva, Stuchi, Violato & Cuozzo, 2017), outperforming the state-of-the-art VADs. Similarly, the magnitude of the audio spectrogram is

employed as input feature in (Tashev & Mirsamadi, 2016), where an MLP-based VAD is proposed.

The issue of localizing a speaker in a binaural context has been addressed in (Ma, May & Brown, 2017), where an MLP predicts the speaker azimuth with the help of a simulated head movement of the listener. The more general sound localization task is performed in (Kovandžić, Nikolić, Al-Noori, Ćirić & Simonović, 2017). In this work the TDOAs measured from signal captured by eight different microphones are used to feed an MLP; after that, the neural network is capable of accurately locate the sound source in near field condition. The model proposed in (Chakrabarty & Habets, 2017) performs speaker localization in terms of azimuth by means of CNNs, by proposing a novel technique for exploiting the phase of audio signals recorded by a linear array. Similarly, CNNs have been employed in (Ferguson, Williams & Jin, 2018) to perform sound source localization. Multiple speakers localization is addressed in (He, Motlicek & Odobez, 2018), where a robot predicts the speaker azimuth in a indoor environment by using a CNN fed with Mel-dependent GCC-PHAT features. The authors exploited CNNs for predicting the speaker coordinates inside the room in multi-room environment in (Vesperini, Vecchiotti, Principi, Squartini & Piazza, 2018), outperforming the state-of-the-art localization algorithm.

Although promising results have been achieved with the new DNN-based VAD and SLOC algorithms, few works target the development of a reliable framework performing speaker detection and localization at the same time. To solve this task, two main strategies can be followed. The first one relies on cooperative but distinct VAD and SLOC algorithms (Tachioka, Narita, Watanabe & Le Roux, 2014; May, van de Par & Kohlrausch, 2012; Chakrabarty & Habets, 2017; Valenzise, Gerosa, Tagliasacchi, Antonacci & Sarti, 2007), while the second one uses one unique model acting simultaneously as detector and localizer (Taghizadeh, Garner, Bourlard, Abutalebi & Asaei, 2011; Vecchiotti, Principi, Squartini & Piazza, 2018). Last but not least, the latest proposed frameworks simultaneously accomplishing VAD and SLOC, rarely make use of new DNN approaches.

In terms of frameworks counting on distinct VAD and SLOC algorithms, a binaural model for speaker detection, localization and recognition is presented in (May, van de Par & Kohlrausch, 2012). This work localizes the speaker by means of a GMM classifier elaborating gammatone filters dependent binaural features. Subsequently, a speech detection module applies a binary mask to GMM azimuth predictions. Similarly, in (Chakrabarty & Habets, 2017) a microphone array beamforming technique divides the room under study into a fixed number of cells, from which features are extracted and classification takes place by means of a GMM. For the audio surveillance purpose, in (Valenzise, Gerosa, Tagliasacchi, Antonacci & Sarti, 2007) a first detection stage is employed, where two separated GMMs classify scream and gunshot signals, respectively. Then localization is performed by means of cross-correlation based TDOA estimation. An ensemble of SLOC and VAD algorithms is studied in (Tachioka, Narita, Watanabe & Le Roux, 2014). This study focuses on the interaction of several classical VAD and SLOC models. Furthermore, an inte-

gration architecture based on DNN or GMM is there proposed, leading to a higher overall accuracy.

With regards to unique VAD and SLOC models, a modified version of the SRP-PHAT is proposed in (Taghizadeh, Garner, Bourlard, Abutalebi & Asaei, 2011), where the SRP-PHAT algorithm processes both speech and noise data, and a rest position is predicted when the latter occurs. The authors discuss a CNN-based model for joint detection and localization for multi-room context in (Vecchiotti, Principi, Squartini & Piazza, 2018). The proposed model exploits localization and detection features, and is able to predict the speech presence and the speaker coordinates by means of multiple outputs. A more accurate localization performance is then achieved in (Vecchiotti, Principi, Squartini & Piazza, 2018) by cascading a CNN-based SLOC trained on true speech data.

1.1. Problem Statement and Motivation

SLOC algorithms proposed in literature so far are generally evaluated within the condition of a perfectly detected speaker activity, or, in other words, in presence of an Oracle VAD. However, this perfect working condition is not satisfied in real-world applications, where VAD systems do commit errors which affect the accuracy of localization algorithms. For this reason, it is necessary to consider the detection and the localization of a speaker in a real-scenario as a unique problem, so that the dependency between VAD and SLOC algorithms can be properly addressed. Hence, this work proposes a novel data-driven framework capable of detecting and localizing a speaker, aiming to limit errors performed by VADs and to increase the overall accuracy of the overall framework.

In particular, this work is an extension of the previous authors' work (Vecchiotti, Principi, Squartini & Piazza, 2018), where an unique CNN-based system for VAD and SLOC was proposed, with purpose to increase the overall accuracy of the two addressed tasks. In (Vecchiotti, Principi, Squartini & Piazza, 2018) it has been observed that the most performing architecture is composed of a Neural SLOC cascaded to the proposed neural model employed as VAD. For this reason, the solution proposed here for the joint detection and localization of a speaker follows the results achieved in (Vecchiotti, Principi, Squartini & Piazza, 2018). In details, the architecture adopted here relies on a first neural VAD based on CNNs, to whom a neural SLOC is cascaded. In particular, within this work several neural VAD models are developed and tested, in order to highlight the advantages of the proposed neural architectures, plus a novel neural SLOC is introduced. The idea behind this approach is to maximise the accuracy and the reliability of a data-driven VAD, so that the minimum amount of wrongly detected speech by VAD is then passed to the following SLOC configured as cascade.

A multi-room environment is considered for evaluating the performance of the proposed framework. Indeed, comparing a single-room scenario and multi-room one, they undoubtedly share some common aspects, however the latter can be considered closer to a real-world application. In particular, both scenarios are subjected to crosstalk between multiple speakers, however in the multi-room environment this event could occur between speakers located in different

rooms. Hence, a model for speaker detection and localization must be robust against utterances pronounced in room different from the one under observation. A similar issue raises for background noise. Indeed, even noise coming from other rooms must be rejected by the VAD and SLOC framework. Last but not least, room-dependent reverberations affect signals in different manners. In conclusion, considering a real world application where noise and speech signals are present inside and outside the room under study, a multi-room scenario succeeds in replicating this working condition.

In addition, a contribution of this work is to highlight the superiority in terms of VAD and SLOC of a framework based on data-driven models against another solution based on classical algorithms. For this reason, the only framework present in literature for joint VAD and SLOC in a multi-room environment is considered for comparison (Tachioka, Narita, Watanabe & Le Roux, 2014). It relies on an ensemble of the state-of-the-art classical VAD and SLOC algorithms, plus an integration stage. The multi-room scenario addressed in this work consists in the Simulated subset of the DIRHA dataset (Cristoforetti, Ravanelli, Omologo, Sosi, Abad, Hagmüller & Maragos, 2014). Furthermore, in order to propose the most reliable data-driven solution, several neural models are developed and compared in this work. In particular, four CNN-based VADs are discussed, which differ in terms of employed input data and neural network outputs. After that, two CNN architectures are proposed for SLOC, and the most performing one is cascaded to the most accurate VAD model.

In order to perform a fair comparison with respect to the baseline method (Tachioka, Narita, Watanabe & Le Roux, 2014), the same simulation strategies are here adopted, which differs from the ones of (Vecchiotti, Principi, Squartini & Piazza, 2018). In details, another version of the DIRHA dataset is here employed, being characterized by two parts of equal length. As a consequence of that, it follows that the cross-validation testing strategy adopted by the authors in (Vecchiotti, Principi, Squartini & Piazza, 2018) is not considered in this work. In addition, simulations in (Tachioka, Narita, Watanabe & Le Roux, 2014) take place at a lower frame rate from what employed by the authors in (Vecchiotti, Principi, Squartini & Piazza, 2018). Here the lowest of the two frame rate is adopted, which corresponds to the one of (Tachioka, Narita, Watanabe & Le Roux, 2014). As result, less data is presented during the model training with respect to (Vecchiotti, Principi, Squartini & Piazza, 2018), and, to address this issue, data augmentation technique is here taken into account. Indeed, the version of the DIRHA dataset previously employed in (Vecchiotti, Principi, Squartini & Piazza, 2018) initially extends the training data. Hence, a specific data augmentation approach is here presented and tested. The idea behind this strategy aims to develop an additional version of the DIRHA dataset, where new speech data is employed. Nevertheless, room impulse responses (RIRs) originally recorded within the DIRHA project are not publicly available, thus a hybrid approach is necessary. The authors decide to virtually emulate the RIRs related to the rooms under study by means of a RIRs generator tool, relying on the only knowledge of the room dimension and the placement of the microphones installations.

This paper first describes the proposed method in Section 2, where the CNN-based VAD and SLOC models are presented. The baseline method used for comparison is then discussed in Section 3. In Section 4 the description of the experimental setup used for voice activity detection and speaker localization is given. Finally results are reported in Section 5, followed by conclusion in Section 6.

2. Proposed Method

The method proposed in this work is depicted in Fig. 1. In details, the detection and the localization of a speaker are performed by means of two distinct algorithms disposed in a cascade configuration. Indeed, speech activity is predicted by the VAD algorithm elaborating audio features extracted from audio signals captured in the room under observation. After that, localization is performed by the SLOC algorithm over speech frames correctly detected by the VAD algorithm. A feature extraction stage precedes the VAD and SLOC models, leading to LogMel and GCC-PHAT Pattern features which feed the proposed neural networks, depending on the models configuration. A simple post-processing technique is employed only for localization predictions. Four different data-driven models for VAD are described in Section 2.2. In particular, they are the Joint-V VAD model proposed in (Vecchiotti, Principi, Squartini & Piazza, 2018); the Joint-S VAD, which shares the same neural architecture of the Joint-V VAD, but performs speech detection by means of different neural outputs; an alternative version of the Joint-V VAD without two of its three outputs, referred to as Alternative Joint VAD (Alt Joint VAD); a simple Neural VAD using input features commonly employed for VAD. After that, two neural architectures trained by means of an Oracle VAD are investigated in Section 2.3 for SLOC, whose architectures allow to differently exploit input data captured by multiple microphones. Localization is performed in terms of speaker coordinates, where the height of the speaker from the ground is not taken into account. Hence, considering the 2-D top view of a room, the speaker Cartesian coordinates will be referred as χ and ψ, being normalized to the [0,1] range by dividing for the wall length.

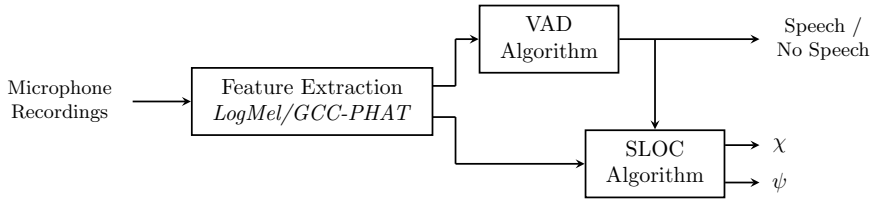

Figure 1: Conceptual scheme of the proposed method. Audio features are extracted from the captured signals, which are used by VAD and SLOC algorithm depending on their specific configuration. After that, the SLOC algorithm performs localization over speech frames detected by the VAD algorithm.

2.1. Features Extraction

Two different features are used within the proposed framework: LogMel and GCC-PHAT Patterns. The first are commonly employed for audio analysis, while the latter are specific for the localization task. Their reliability has been assessed in authors' previous contribution (Vecchiotti, Principi, Squartini & Piazza, 2018).

2.1.1. LogMel

The extraction process of LogMel (Davis & Mermelstein, 1980) is as follows: the audio signal is divided in partially overlapped frames. Then the Discrete Fourier transform is computed, and a set of filters uniformly spaced in the mel-frequency scale is applied in the frequency domain. LogMels are finally obtained by calculating the energy in each sub-band and taking its logarithm. In this case study a set of 40 Mel filters is considered, while the signal is framed with hop size and frame size equal to 50 ms and 60 ms respectively. LogMel features go through zero mean and unit variance normalization.

2.1.2. GCC-PHAT Patterns

The purpose of this feature (Knapp & Carter, 1976) is to estimate the delay between two audio signals recorded by a microphone pair in the presence of the same sound event. Indeed, due to sound propagation, the sound wave reaches the two microphones in different time instants. This behaviour allows to estimate the Direction of Arrival (DOA) of the audio event. GCC-PHAT Patterns computation relies on the frequency domain cross-correlation between the two microphones audio signals, from which the Fourier inverse transform is then applied. Only the first 51 values of the inverse transform are selected, since adjacent microphones pairs distancing 50 cm are considered for feature extraction. Frame size and hop size of 50 ms and 60 ms respectively are used in the features extraction stage. Finally, features are normalized in the range [0, 1].

2.2. Voice Activity Detection

In this work four neural models for the VAD task are discussed and compared. The first one is the Joint-V VAD model previously proposed in (Vecchiotti, Principi, Squartini & Piazza, 2018), where it was referred to as Joint VAD-SLOC. In details, the name *Joint* stays for the employment of both detection and localization features, *-V* is for the use of its detection output, while *VAD* means that the model is employed for speech detection. After that, the Joint-S VAD model is reported. It does not differ from the Joint-V VAD, except from how the speaker activity is determined. Indeed, this model makes use of its localization output instead of the VAD output to estimate the presence of speech. Finally, the description of the Alt Joint VAD model and of the Neural VAD model is given.

2.2.1. Joint-V VAD

This model addresses both the localization and detection tasks in a multi-task learning framework (Caruana, 1997). Indeed, the authors have previously shown (Vecchiotti, Principi, Squartini & Piazza, 2018) that using both the localization and detection information improves the VAD accuracy. The Joint-V VAD is depicted in Fig. 2. It consists of a CNN fed by LogMel and GCC-PHAT Patterns features, and it is trained by means of three outputs dedicated to both speech detection and speaker localization. Two different branches of convolutional layers processes the two features sets, then a concatenation of the branch-dependent feature maps is performed. These two branches have the same neural architecture, or, in other words, share the same hyper-parameters. After that a set of hidden neuron layers is applied. The model ends with three outputs, where the first one estimates the speech presence, and the remaining two correspond to the speaker coordinates inside the room in a 2-D plane.

This model jointly acts as detector and localizer. The speech detection dedicated output makes use of labels assuming 0 or 1 value, while the localization task is treated as a regression problem, hence the two localization outputs are mapped in the continuous [-1, 1] range. In details, when speech is present, the speaker is given in the range [0, 1] for both coordinates, while both labels are set to -1 in the case of speaker inactivity. Following this approach, both the detection output and the coordinates outputs are valid for speech detection. Indeed, in (Vecchiotti, Principi, Squartini & Piazza, 2018), speech detection was performed by means of the localization outputs of the network, by applying a linear threshold located in a 2-D plane. In this work, speech detection is performed only by means of the single detection output, so that a possible confusion regarding the employment of the network outputs is avoided. In addition, due to the new splitting strategy for the considered dataset, less data is available for training the model. Hence, the possibility that insufficient speech data will be presented during the model training must be taken into account. As a consequence, it is reasonable to expect that this condition affects more the localization outputs than the speech detection output, since the latter is just a binary label.

Due to the [-1, 1] range, *hard tanh* is employed as activation function of the localization outputs, while *sigmoid* activation is used for the detection output. A temporal context extends the amount of data processed by the network frame-by-frame. With this procedure, previous and future frames are processed together with the actual frame, for a total of C frames, where C denotes the *context*. Consistently with authors' previous work (Vecchiotti, Principi, Squartini & Piazza, 2018), the selection of past and future frames is piloted by the integer value *strides*, although this value is here set equal to 1. In details, a 2-D matrix is obtained for each microphone for the actual frame, where the rows are the features and the columns are the frames with context (Fig. 2). Then the different microphones features are stacked, leading to a 3-D tensor. The model training is performed on speech and non speech data.

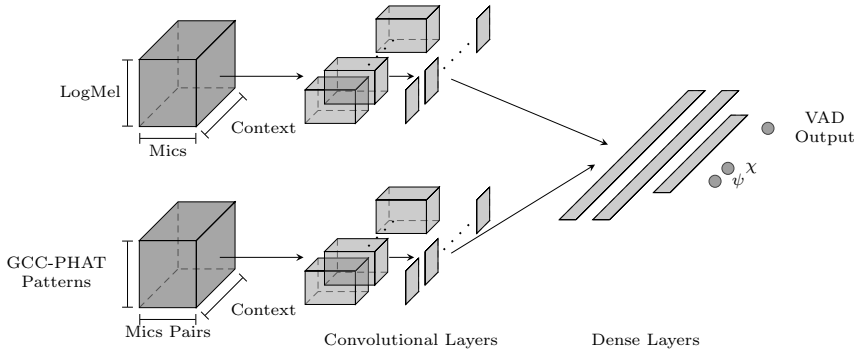

Figure 2: Architecture of the Joint-V VAD model. Two separated convolutional layers elaborate LogMel and GCC-PHAT Patterns features, respectively. Since the input matrices are 3-D, the first convolutional layer has 3-D kernels, which then become 2-D. A concatenation step joins feature maps extracted by the two convolutional stacks.

2.2.2. Joint-S VAD

This model shares the same neural architecture with the Joint-V VAD, making use of detection and localization feature, and being characterized by three outputs (Fig. 2). However, the speaker activity is determined by means of the localization outputs instead of the detection one, since these two outputs are eligible for VAD, as discussed in the author's previous work (Vecchiotti, Principi, Squartini & Piazza, 2018). Speech detection is then performed by means of a particular threshold, which corresponds to a oblique line in the 2-D plane of the room. The purpose of this model is to properly compare the Joint-V VAD, plus to show that also its SLOC outputs can be accurately trained, even if their training is more sensible to employed data compared to its VAD output.

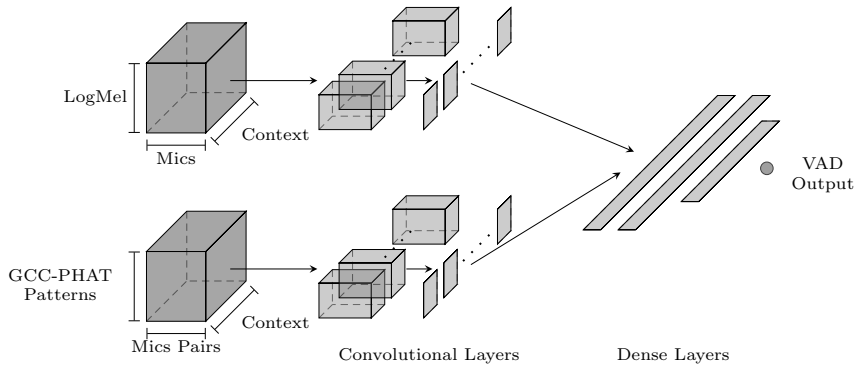

Figure 3: The Alt Joint VAD model. Its architecture shares many aspects with the Joint-V VAD shown in Fig. 2, however the χ and ψ outputs are absent. This model is used for comparison, aiming to show the importance of using the speaker coordinates for the network training.

2.2.3. Alt Joint VAD

This model shares many aspects with the Joint-V VAD, and it is depicted in Fig. 3. The key difference is the absence of the two outputs dedicated to localization. The purpose of this model is to directly compare the Joint-V VAD, in order to highlight the importance of the two SLOC outputs for the model training, even if they are not evaluated in terms of speaker localization accuracy. In addition, since the single output of the model is not comprised in the range [-1,1], ReLU is employed as activation function instead of hard tanh.

2.2.4. Neural VAD

This neural architecture for VAD has been already addressed in (Vecchiotti, Principi, Squartini & Piazza, 2018). Its block diagram is shown in Fig. 4, it processes only detection features, and no SLOC outputs are present at the end of the network. The aim of this model is to show the importance of localization features for the detection task. Even in this case, ReLU is employed as activation function.

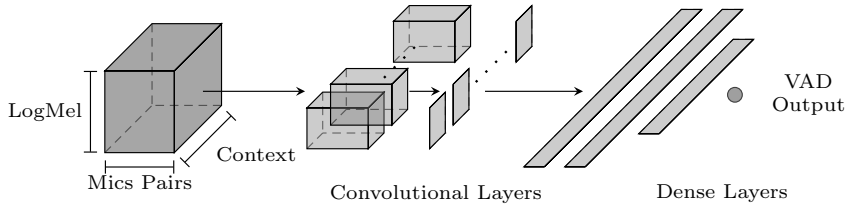

Figure 4: The Neural VAD model. The purpose of this model is to compare the proposed Joint-V VAD model with a standard CNN-based VAD which is the result of the previous authors' contribution (Vecchiotti, Principi, Squartini & Piazza, 2018).

2.3. Speaker Localization

Encouraging results have been achieved by employing DNNs (Ma, May & Brown, 2017) and especially CNNs (Chakrabarty & Habets, 2017; Vesperini, Vecchiotti, Principi, Squartini & Piazza, 2018) for this task.

In this work, the authors use two CNNs architectures to perform localization. Both networks are trained on speech data by means of an oracle VAD, and their outputs are the room coordinates in the range [0,1]. *ReLU* is selected as activation function.

The first model is the same discussed in (Vecchiotti, Principi, Squartini & Piazza, 2018), and it is referred to as Single-Channel SLOC (SLOC_{SC}). The GCC-PHAT Patterns feature are organized in a 3-D tensor, as discussed in Section 2.2. The second model differs from the previous one in terms of input features organization and elaboration. Indeed, a standalone input is created for each pair of microphones. As result, a set of 2-D matrices is now presented to the network, where rows and columns of each matrix are the temporal context and the features. The CNN is then characterized by a number of inputs equal to the considered microphone pairs. The name adopted for the model is SLOC_{MC}

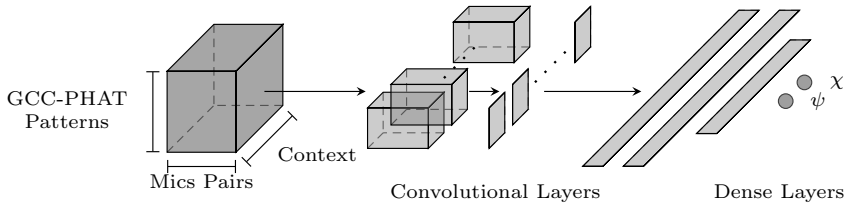

Figure 5: Single-Channel architecture.

(Multi-Channel SLOC). The main difference between the two models is how the first CNN layer processes the inputs. Indeed, in the case of the SLOC_{SC}, the first CNN layer consists in a set of 3-D kernels. For each kernel a 2-D feature map is then computed, where a summation over the third dimension takes place. In details, this summation acts as a compression stage over the extracted microphone-dependent feature maps. Differently, in the latter case, the first CNN layer consists in 2-D kernels, which are trained over data coming from different microphones, and no feature maps compression is performed.

Finally, the SLOC output is further processed by using a smoothing technique. In details a moving average filter of window size equal to 5 is applied to each predicted coordinate.

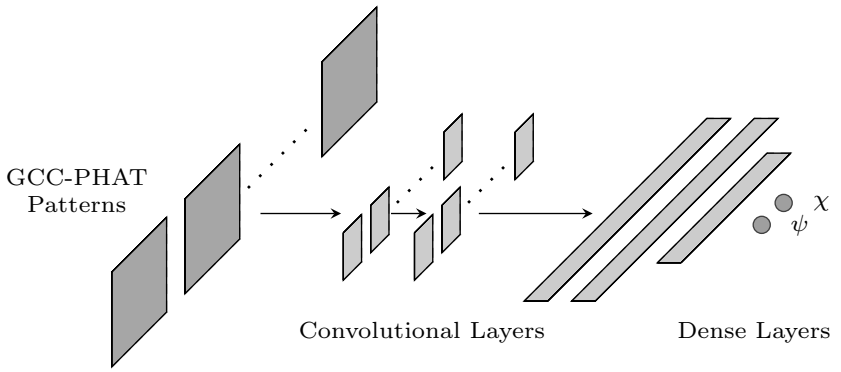

Figure 6: Multi-Channel architecture.

2.4. Data Augmentation

A specific data augmentation technique is employed in this work. Indeed, in (Krizhevsky, Sutskever & Hinton, 2012) it has been observed that the accuracy of a data-driven algorithm improves when extra-data is used for the model training, since chances of overfitting the model are reduced. Data augmentation was initially applied for image recognition purpose (LeCun, Bottou, Bengio & Haffner, 1998), where images already present in the dataset were processed, in order to generate new training material. The effectiveness of data augmentation

has been recently assessed even in the audio field. In particular, in (Tüske, Golik, Nolden, Schlüter & Ney, 2014) data augmentation allows to achieve a higher accuracy when applied for the language recognition purpose. Similarly, this technique has found application also in speech recognition (Cui, Goel & Kingsbury, 2015; Ragni, Knill, Rath & Gales, 2014; Schlüter & Grill, 2015; Prisyach, Mendelev & Ubskiy, 2016; Zhou, Xiong & Socher, 2017; Ko, Peddinti, Povey, Seltzer & Khudanpur, 2017) and sound event detection (Tran, Ng & Leng, 2017; Zöhrer & Pernkopf, 2017; Salamon & Bello, 2017).

In this work, data augmentation targets VAD and SLOC; furthermore, up to the authors' best knowledge, data augmentation has never been adopted for the sound localization task. Two main strategies are here employed. The first one is an extension of the dataset by means of external data already recorded in the same conditions of the dataset under study (Cristoforetti, Ravanelli, Omologo, Sosi, Abad, Hagmüller & Maragos, 2014). The other approach requires to generate a new dataset from scratch. The proposed data augmentation technique generates virtual acoustic scenes using appropriate audio software and some parameters of the real scene. Further details are then reported in Section 4.2. As result, this second technique is suitable to be applied to different case studies, being independent from the dataset taken into account.

3. Baseline Method

A brief description of the baseline model proposed in (Tachioka, Narita, Watanabe & Le Roux, 2014) and employed here for comparison is reported in this section. The baseline model consists in an ensemble of multiple VAD and SLOC algorithms. Indeed, in (Tachioka, Narita, Watanabe & Le Roux, 2014) two algorithms are considered for VAD, being Sohn's method and Switching Kalman Filter (SKF). Four SLOCs algorithms are evaluated, where three are derived from the Cross Spectrum Phase method, being 2D-CSP, multi-channel CSP and Template CSP, and the last SLOC algorithm is Steered Response Power (SRP-PHAT). A final integration algorithm jointly processes VAD and SLOC predictions. Three methods are investigated: Minimum Cost Criterion, Support Vector Machine (SVM) and a neural network based classifier. A three stages selection strategy leads to the best configuration of this ensemble, which is the one reported in this work.

3.1. Voice Activity Detection

The first of the two detection algorithm is Sohn's method (Sohn, Kim & Sung, 1999), which is based on conventional likelihood ratio test. This method assumes that the noise power spectra estimated in speech frames is conditionally independent from its observation in non-speech frames. Hence the statistical models of speech and noise are formulated, being characterized by their variance, respectively. For each frequency bin the log-likelihood ratio of the speech and noise models is computed, and the geometric mean is then computed. Finally, a threshold is applied to this mean in order to determine the class of the frame under observation.

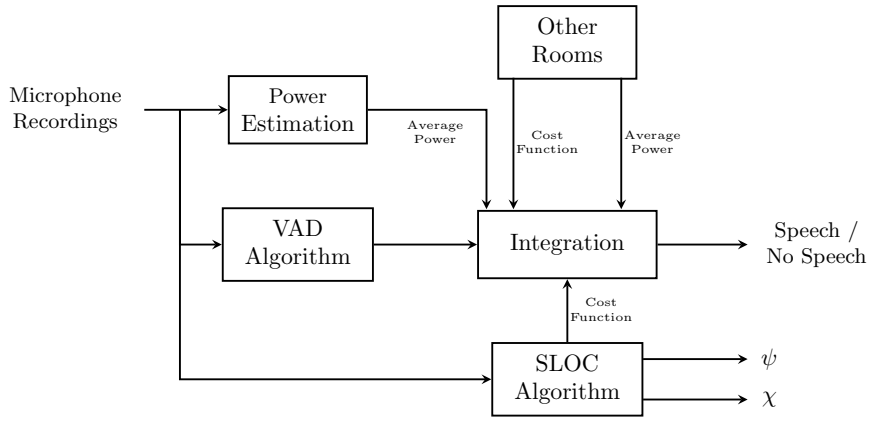

Figure 7: Conceptual scheme of the baseline method

The other detection algorithm is switching Kalman filter (Fujimoto & Ishizuka, 2008). It relies on a prepared speech model and an on-line estimated noise model, from which the noisy speech model is finally build. This approach elaborates LogMel features by means of a GMM which is continuously updated by the Kalman filters. The model ends with a likelihood ratio, which allows to discriminate speech and noise.

3.2. Speaker Localization

A modified version of the original CSP method (Knapp & Carter, 1976) is chosen used for speaker localization. Indeed, the original method assumes a plane sound wave, while in the reference work the 2D-CSP is tested under the spherical wave assumption. This method estimates the Time Difference of Arrival (TDOA) between two adjacent microphones. For this purpose the frequency domain cross-correlation of the two signal is computed, similarly to what described in Section 2.1 for GCC-PHAT Patterns. From the cross-correlation the maximum of the inverse transform is taken. After that, for a speaker candidate point a cost function is defined, which consists in the difference between the theoretical and the estimated time delays related to each considered microphone pair. The speaker position is eventually given as the point that minimize this cost function.

In addition, in the baseline method (Tachioka, Narita, Watanabe & Le Roux, 2014) the multiple channel 2D-CSP (M-CSP) (Hayashida, Morise & Nishiura, 2010) is taken into account. This technique extends conventional CSP by using a correlation matrix of time difference of arrival.

The third SLOC algorithm discussed in (Tachioka, Narita, Watanabe & Le Roux, 2014) is the Template CSP, which is a modified version of the 2D-CSP. Indeed, since the theoretical TDOAs and the observed TDOAs differ due to reverberation present in the room under observation, the theoretical TDOAs are subjected to a correction. In details, a bias is added to the coordinates of

each position, where the bias is estimated as the average difference between the theoretical and the observed TDOA in the development set.

The last algorithm used for SLOC is the SRP-PHAT (Do, Silverman & Yu, 2007). This technique steers a delay-and-sum beamformer in the volume under observation. From that, an objective function is then computed, which depends on the frequency domain cross-correlation of the signals recorded by microphone pairs. Thus PHAT weighting procedure is applied. Finally, with Stochastic Region Contraction (SRC) the area under observation is recursively reduced. The speaker position is finally estimated as the point maximizing the objective function.

3.3. Integration

The first integration algorithm relies on minimum cost criterion. It is applied when a speaker is detected in multiple rooms. Hence, the localization cost function is compared across the detected rooms, and the smallest one determines the room prediction. However, since the cost function depends on the room size, a tolerance parameter is introduced. This parameter associates a flag to the evaluated frame when the cost function is close to be the smallest between the detected rooms, instead of being the smallest in absolute. Finally, the utterance under study is rejected if the ratio of the flags over the total number of utterance frames is lower than a threshold.

The other two approaches are classifier-based requiring a training stage. In details, features from all the rooms are fed to the classifier, which predicts the probability of having speech only for the room under observation. This prediction is then flagged by means of a tolerance parameter, exactly as for the cost criterion. Similarly, a threshold is applied to the ratio of the recognized frame over the total of utterance frames.

Two different classifiers are tested. Their input features are the speech powers averaged over microphones in each room and the localization function cost. The first classifier is a SVM, while the other one is a MLP consisting of two hidden layers of 15 and 10 units each. A standalone classifier is trained separately for each room.

3.4. Comparison with the Proposed Method

The baseline model is composed of a VAD algorithm, a SLOC algorithm and an integration stage. This model results to be complex for a couple of aspects. First of all, a dedicated manual tuning is required for each one of the VAD, SLOC and integration algorithm, which can be extremely time demanding. Furthemore, each room must be analysed before the single-room prediction. These issues are addressed by the proposed method. Indeed, a extensive tuning for each algorithm is not required and the other room predictions are not necessary when the model is applied to the room under study. In addition, the proposed method avoids a third integration stage.

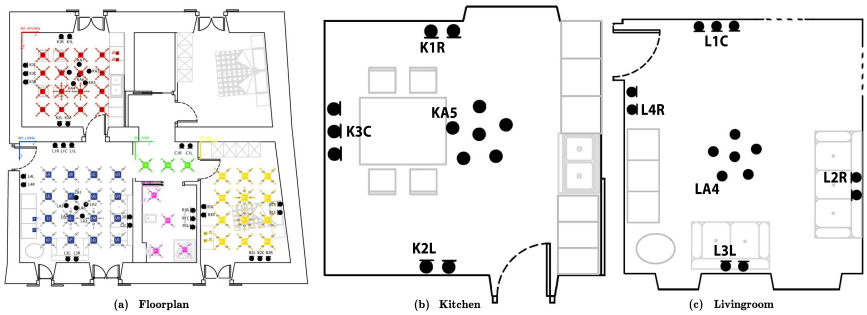

Figure 8: The map of the apartment used for the DIRHA project (a). Figures (b) and (c) show the considered rooms, where the thick black dots are the installed microphones.

4. Experimental Setup

4.1. DIRHA Dataset

The DIRHA project (Cristoforetti, Ravanelli, Omologo, Sosi, Abad, Hagmüller & Maragos, 2014) targets the tasks of speech detection, localization and recognition in a domestic environment. It consists of a set of recordings in a 5 rooms apartment. A total of 40 omnidirectional microphones is installed in the walls and in the ceilings of the apartment, as shown in Fig. 8. Adjacent microphones are spaced by 50 cm. Walls installations measure about 200 cm from the ground, while ceiling installations are present only in the kitchen and living room.

Two subsets split the DIRHA dataset, the *Real* and the *Simulated*. The first one consists in real recordings, with moving speakers, while the second one is obtained by convolving a fixed number of measured RIRs with speech data. In addition, the latter is characterized by overlapping speech events, while this condition is not present in the Real subset. In this work experiments have been performed on the Simulated dataset, since a higher amount of speech is available. Moreover, the proposed methods are tested in the kitchen and living room of the apartment, since a ceiling installation is present and most of the speech events are expected to occur in these two rooms. In details, 17 speaker positions are available for both kitchen both living room, while the latter counts a total of 15 microphones and the kitchen 13 installations.

4.1.1. HSCMA and EVALITA

Two different versions of the Simulated DIRHA dataset are taken into account in this work. The EVALITA dataset has been employed by the authors in their previous contribution (Vecchiotti, Principi, Squartini & Piazza, 2018; Vesperini, Vecchiotti, Principi, Squartini & Piazza, 2018). The Simulated EVALITA contains 80 scenes of Italian spoken utterances. This dataset was used in (Vecchiotti, Principi, Squartini & Piazza, 2018; Vesperini, Vecchiotti, Principi, Squartini & Piazza, 2018) in which the experiments were carried out using the k-fold cross validation technique, where $k = 10$, so that 64-8-8 scenes

compose the training-validation and test sets. On the contrary, the baseline approach (Tachioka, Narita, Watanabe & Le Roux, 2014) is tested over Simulated and Real subset of the HSCMA dataset. This dataset contains 80 samples of one minute length, equally divided in Italian, Greek, German and Portuguese languages. The Simulated HSCMA dataset is divided in the *Dev* and *Test* subsets, each composed of 40 scenes of one minute length; the first is employed for training the model and the second for testing its performance. In details, training and validation sets for CNN training are obtained from the *Dev* set, with a 90% and 10% split, respectively.

4.2. DIRHA-LibriSpeech

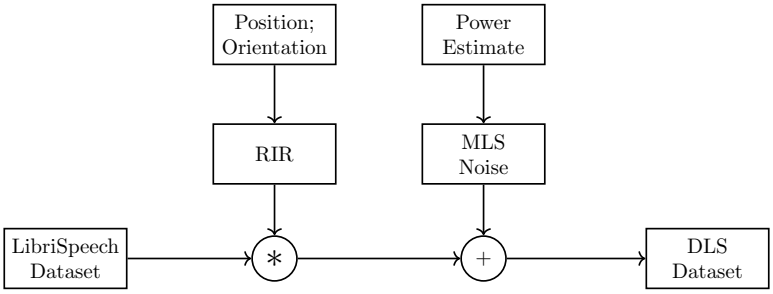

Figure 9: Block diagram of the algorithm used for the realization of the DLS dataset

The proposed method consists also in a data augmentation stage, and the newly created dataset will be denoted as DLS (DIRHA-LibriSpeech). The data augmentation technique being adopted here creates new data by replicating the acoustic scene of the considered rooms. Since the original RIRs recorded within the DIRHA project are not publicly available, a new set of RIRs must be generated consistently. For this purpose, a version of the Python Room Impulse Response generator (Sivasankaran, Vincent & Campbell, 2017) is employed, which relies on the Image Source Model theory (Allen & Berkley, 1979). The artificial dataset aims to replicate the working condition of the DIRHA Simulated subset, where the speaker positions are fixed. The rooms under observation are the kitchen and the living room. In the first case, 17 positions are available for the speaker, where each position can assume 4 different orientations as shown in Fig. 11, in addition 13 microphones are installed in this room. As result, 884 RIRs are computed. For the livingroom, displayed in Fig. 10, the number of speaker positions and orientation is the same as the kitchen, however 15 microphones are installed, leading to 1020 RIRs.

Speech data employed for DLS is randomly selected from the LibriSpeech dataset (Panayotov, Chen, Povey & Khudanpur, 2015). Only the clean speech subset of LibriSpeech is considered for this purpose. A total of 500 utterances from the LibriSpeech daset is employed for the DLS creation. A desired SNR is then achieved by adding artificial noise created with maximum length sequence (MLS) technique. MLS amplitude is calculated for each LibriSpeech utterance.

As result, the same noise power characterizes each microphone. The block scheme of the DLS development is depicted in Fig. 9.

Important differences occur between DLS and DIRHA EVALITA or HSCMA. At first place, the latter is the result of the measured RIRs between the positions of the sources (using acoustic loudspeakers) and the microphones installations, while the DLS is the result of the modelled RIRs. In addition, the language employed for the DLS is English, while EVALITA is in Italian and HSCMA has speech pronounced in Italian, Greek, Portuguese and German. Finally, the DIRHA project contains scenes and overlapping events coming from other rooms, while the authors decide to avoid this option for the DLS.

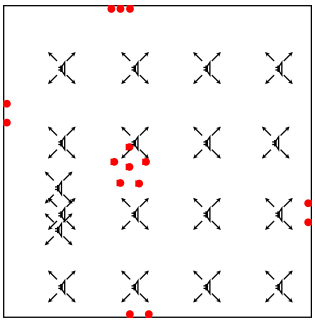

Figure 10: The living room designed through the data augmentation process.

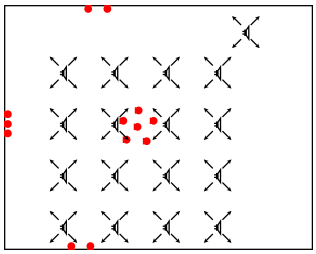

Figure 11: The kitchen designed through the data augmentation process.

4.3. Evaluation Metrics

Two dedicated groups of metrics assess the performance of VAD and SLOC algorithms. Moreover, it is necessary to consider the dependency of SLOC performance from the VAD one, since speaker localization is performed after speech detection due to cascade configuration. This issue is here tackled by testing SLOC accuracy over speech frames correctly detected by VAD (true positive). The authors decided to employ the same metrics of their previous contribution (Vecchiotti, Principi, Squartini & Piazza, 2018), in order to have a valid reference for the algorithms deployment. Furthermore, the state-of-the-art work (Tachioka, Narita, Watanabe & Le Roux, 2014) is here tested consistently with the same metrics employed by the authors.

Three metrics evaluate the VAD performance, namely the false alarm rate (FA), the deletion rate (Del) and the overall speech activity detection (SAD) defined as:

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

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 $\beta = N_{nsp}/N_{sp}$ balances the different amount of data between speech and non speech in the test set.

Root Mean Square Error (RMSE) and P_{cor} measure the localization accuracy. RMSE is defined as:

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

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 latter is defined as $P_{cor} = N_{FINE}/N_{TOT}$, where N_{FINE} is the number of frames localized with RMS inferior than 500 mm.

4.4. Neural Networks Details

The GPU-based toolkit *Keras* (Chollet et al., 2015) has been employed for developing and testing the DNN models. Two computers have been exploited for simulations: the first one is an HP notebook model *15-p257nl* equipped with a 4-core Intel i7 2.4 GHz, 16 GB of RAM and a *Nvidia GeForce 840M* graphic card; the second one is equipped with a 6-core Intel i7, 32 GB of RAM and a *GeForce GTX970* graphic card.

Training and testing of the proposed models rely on two different susbsets of the DIRHA Simulated dataset. In details, *Dev* subset is used for training the DNNs and for optimizing the hyper-parameters of the baseline model. Testing is executed over the *Test* subset. When data augmentation is used, the *Dev* training set is extended with new data, while *Test* is not varied.

Differently from the authors' previous work (Vesperini, Vecchiotti, Principi, Squartini & Piazza, 2018), here microphone selection is not considered. Indeed, when LogMel features are extracted, all the available microphones are taken into account, being in total 13 and 15 for the kitchen and the living room, respectively. GCC-PHAT Patterns are extracted from adjacent microphone pairs. In details, with regards to the ceiling array, all the possible combinations have been considered, for a total of 15. Hence, a total of 19 and 20 microphone pairs is selected for the kitchen and the living room, respectively.

The DNN optimization strategy here adopted relies on two stages. In the first stage, the neural network architecture is investigated by means of a random search technique; after that, the most performing model resulting from the previous stage is trained again by using the augmented dataset.

Training of the neural networks is performed by means of *Adam* optimizer, of which decay parameter for momentum estimates are $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The number of training epochs is set to 500, while a batch size of 200 frames is employed. Neural network weights are initialized with a gaussian distribution with $\mu = 0$ and $\sigma = 0.1$. In addition, strides are used in convolutional kernels, in order to let the neural network process the equivalent of a larger audio excerpt. Furthermore, convolutional kernels go through two regularizers, which take care of the activity of a kernel and the weights decay. Their coefficients *L1* and *L2* are both set to 1e−4. The Joint-V VAD, Joint-S VAD, Alt Joint VAD and Neural VAD are trained with learning rate equal to 5e−5, while the two SLOC models are trained with this value set to 1e−4. Finally, overfitting is prevented

by applying early stopping after 5 epochs without improvement on the validation loss. Variable learning rate allows a finer tuning of the models, by decreasing of a factor scale 0.5 after 2 epochs without improvements. The authors decided to set the *context* value to 15 frames for both models, since this parameter has been deeply investigated in their previous work (Vesperini, Vecchiotti, Principi, Squartini & Piazza, 2018). Dropout equal to 0.5 is applied after each hidden layer. The investigated hyper-parameters by random search are reported in Table 1.

5. Results

5.1. Voice Activity Detection Algorithms

Within this section the results achieved by the four proposed models for VAD are compared and discussed, with the purpose of finding the most reliable data-driven solution. Performance of the Joint-V VAD, Joint-S VAD, Alt Joint VAD and Neural VAD models during the two optimization stages are reported in Table 2, respectively.

Initially, the CNN architectures have been investigated by means of random search. In details, the most accurate Joint-V VAD is composed one CNN layer with 128 kernels of shape [5, 5] and strides [4, 4], followed by four hidden layer of 256, 512, 2048, 1024 neurons respectively. The optimized Joint-S VAD architecture is the same of the Joint-V VAD. The Alt Joint VAD achieving the best performance is composed of one CNN layer with 128 kernels with shape [3, 3], strides [2, 2] and four hidden layers with 2048, 1024, 1024, 1024 neurons respectively. Finally, the most performing Neural VAD is composed of two CNN layers with 128 and 64 kernels respectively, having kernels of shape [5, 5] and [3, 3] for each layer and strides [1, 1] and [3, 3], followed by four hidden layers

		Joint-V VAD Joint-S VAD Alt Joint VAD Neural VAD	$SLOC_{SC}$ $SLOC_{MC}$
Convolutional Layers	Number of Layers	1, 2	1, 2
	Number of Kernels	64, 128	64, 128, 256
	Kernel Size	3, 4, 5	3, 4, 5
	Kernel Strides	1, 2, 3, 4, 5	1, 2, 3, 4, 5
Hidden Layers	Number	1, 2, 3, 4	1, 2, 3, 4, 5, 6, 7
	Neurons	256, 512, 1024, 2048	512, 1024, 2048

Table 1: Hyper-parameters of the DNN models, investigated through random search in the first optimization stage.

composed by 512, 512, 256, 512 neurons respectively. At this stage, the most performing model is the Joint-V VAD, which achieves the lowest average SAD of 8.3% over the *Test* subset.

After that, the second optimization stage takes place, where data augmentation is performed. To distinguish this stage the symbol † is appended to the name of the considered model. As expected, all the four models benefit from data augmentation. In particular, the proposed Joint-V VAD model achieves the lowest average SAD of 3.7%, consistently with (Vecchiotti, Principi, Squartini & Piazza, 2018). Furthermore, the Alt Joint VAD model performs worse than the Joint-V VAD. Thus, we can say that the presence of the two SLOC outputs helps during the model training, reason why the Joint-V VAD outperforms the Alt Joint VAD. In addition, the Alt Joint VAD performs worse than the Neural VAD, confirming that the SLOC outputs are extremely important in order to use localization data. On the other hand, the Joint-S VAD achieves the highest SAD also when data augmentation is considered. This behaviour shows that the training procedure of the localization outputs of the proposed joint model is extremely sensible to employed data. Furthermore, it has been previously observed in (Vecchiotti, Principi, Squartini & Piazza, 2018) that these two outputs acting as localizers are less accurate than a Neural SLOC. For this reason as well, a strategy relying on two distinct VAD and SLOC algorithms is chosen in this work.

5.2. Speaker Localization Algorithms

To test the CNN-based SLOCs two strategies are here adopted: the first one couples the SLOC with an Oracle VAD, reported in Table 3, while the latter tests the SLOC over true positive frames detected by the Joint-V VAD†, shown in Table 4. These two strategies are generally consistent in terms of performance, however the first assesses the localization algorithm in an absolute sense, while the second one considers the dependency from the VAD algorithm.

In addition, even the Joint-S VAD model is eligible to be evaluated in terms of localization, as in (Vecchiotti, Principi, Squartini & Piazza, 2018). However the authors decide to not report these results within this section, since this model performs generally worse compared to the SLOC_{SC} and SLOC_{MC}, consistently with (Vecchiotti, Principi, Squartini & Piazza, 2018).

The first optimization stage, where CNNs parameters are varied, leads to the best result of 747 mm RMS achieved by SLOC_{MC} tested with an Oracle VAD. In details, this model has one convolutional layer with 128 kernels each sized 3×3, making use of strides value equal to 2. After that five hidden layers of 512, 256, 2048, 1024, 1024 neurons end the network. Similarly, SLOC_{SC} achieves an almost equal RMS. It is composed of two convolutional layers where a total of 64 kernels of shape [5, 5] and [3, 3] with strides equal to [4, 4] and [3, 3] are employed. The convolutional layer is then followed by four fully connected layers counting 2048, 1024, 1024, 2048 units each. Subsequently, the two best models are trained by using the augmented training set. This step is denoted with †, consistently with Section 5.1. As result, the $\text{SLOC}_{\text{MC}}{}^\dagger$ achieves an

		Kitchen	Living Room	Average
Joint-V VAD	SAD (%)	7.6	9.0	8.3
	Del (%)	9.3	16.3	12.8
	FA (%)	5.9	1.7	3.8
Joint-V VAD†	SAD (%)	4.7	2.7	**3.7**
	Del (%)	7.4	3.5	5.4
	FA (%)	2.0	1.9	1.9
Joint-S VAD	SAD (%)	9.9	11.3	10.6
	Del (%)	16.9	21.5	19.2
	FA (%)	3.0	10.5	6.7
Joint-S VAD†	SAD (%)	7.2	8.6	7.9
	Del (%)	13.7	16.9	15.3
	FA (%)	0.7	0.3	0.5
Alt Joint VAD	SAD (%)	8.2	8.9	8.6
	Del (%)	13.9	15.9	15.0
	FA (%)	2.5	1.9	2.2
Alt Joint VAD†	SAD (%)	6.1	3.7	4.9
	Del (%)	11.4	6.7	9.0
	FA (%)	0.7	0.7	0.7
Neural VAD	SAD (%)	8.4	11.3	9.9
	Del (%)	8.8	16.5	12.6
	FA (%)	8.1	6.2	7.1
Neural VAD†	SAD (%)	4.6	3.9	4.3
	Del (%)	5.9	5.4	5.7
	FA (%)	3.2	2.7	2.9

Table 2: Achieved results for the three proposed data-driven algorithms on the *test* set. For each model the first main line corresponds to the first optimization stage, where neural networks hyper-parameters are investigated. The second line shows the result when data augmentation is applied, denoted with †.

Oracle VAD		Kitchen	Living Room	Average
SLOC_{SC}	RMS (mm)	757	745	751
	P_{cor} (%)	62.8	63.2	63.0
$\text{SLOC}_{\text{SC}}{}^{\dagger}$	RMS (mm)	508	436	472
	P_{cor} (%)	85.8	90.8	88.3
SLOC_{MC}	RMS (mm)	788	707	747
	P_{cor} (%)	57.5	66.7	62.1
$\text{SLOC}_{\text{MC}}{}^{\dagger}$	RMS (mm)	447	415	**431**
	P_{cor} (%)	90.4	94.0	92.2

Table 3: Results for the two proposed SLOC when tested in the presence of an Oracle VAD detecting speech over the *Test* subset. The † denotes the application of data augmentation.

RMS of 431 mm, while a slightly worse performance of 472 mm characterizes

the $\text{SLOC}_{\text{SC}}{}^{\dagger}$.

Joint-V VAD†		Kitchen	Living Room	Average
SLOC_{SC}	RMS (mm)	724	600	662
	P_{cor} (%)	66.5	69.3	67.9
$\text{SLOC}_{\text{SC}}{}^{\dagger}$	RMS (mm)	451	399	425
	P_{cor} (%)	87.8	91.3	90.0
SLOC_{MC}	RMS (mm)	745	563	654
	P_{cor} (%)	61.1	74.6	67.8
$\text{SLOC}_{\text{MC}}{}^{\dagger}$	RMS (mm)	367	377	**372**
	P_{cor} (%)	93.0	95.3	94.1

Table 4: Performance of the two VADs when tested over true positive frames detected by the Joint-V VAD†.

When tested in the presence of the Joint-V VAD† as reported in Table 4, the $\text{SLOC}_{\text{MC}}{}^{\dagger}$ and the $\text{SLOC}_{\text{SC}}{}^{\dagger}$ achieve 372 mm and 425 mm of RMS, respectively. A maximum P_{cor} of 94.1% distinguishes the $\text{SLOC}_{\text{MC}}{}^{\dagger}$. Hence, it is possible to state that the novel SLOC_{MC} architecture is capable of better exploiting data recorded from multiple microphones, which confirms the authors' idea of providing to the CNN a better capability of generalizing compared to the SLOC_{SC}.

As assessed in the authors' previous work (Vecchiotti, Principi, Squartini & Piazza, 2018), SLOC performance increases in the presence of a real VAD. Indeed, since true positives consist in a subset of all the speech data present in the test set, it is possible to state that a real VAD fails to detect speech being more difficult to localize.

5.3. Comparison with the Baseline Method

In Table 5 the best results achieved in (Tachioka, Narita, Watanabe & Le Roux, 2014) are reported. In details, in the baseline model a three stage optimization strategy has been adopted to select the most performing algorithms within the ensemble. Initially, all the four SLOCs are tested in presence of an Oracle VAD. The two more accurate techniques are then separately coupled with the Sohn's and SKF. In this stage, the less performing of the two previously selected SLOCs is rejected. Finally, the three proposed integration algorithms are applied to the remaining SLOC coupled with the two VADs. As result, the best combination is the Sohn's VAD and the Template method as SLOC, when integration is performed by SVM. Here a straightforward notation is adopted for these algorithms. Indeed, Sohn's method plus the SVM integration will be referred as VAD_{B} (Baseline), and the Template SLOC is referred as SLOC_{B}. Specific results for the kitchen and the living room are not available in (Tachioka, Narita, Watanabe & Le Roux, 2014).

Last but not least, the authors report the result of the SLOC_{B} when it is coupled with an Oracle VAD instead of VAD_{B}. Indeed, this result, being shown

		Average
VAD_B	SAD (%)	6.7
	DeL (%)	6.1
	FA (%)	6.1
$SLOC_B$	RMS (mm)	961
	P_{cor} (%)	59.2

Table 5: Results achieved with the most performing algorithms in the baseline method

Oracle VAD		Average
$SLOC_B$	RMS (mm)	1094
	P_{cor} (%)	56.4

Table 6: Best performance of the baseline SLOC in the presence of an Oracle VAD.

in Table 6, is important in order to analyse the baseline SLOC independently from VAD accuracy.

After that, the overall performances of the proposed approach and the baseline model are discussed. In Table 7 a comparison between the two approaches for speaker localization is presented. In details, for each employed metric, Δ is defined as the subtraction of the result achieved by the baseline model from the result related to the most performing algorithm by the authors. Indeed, the data-driven $SLOC_{MC}{}^{\dagger}$ and the baseline $SLOC_B$ are tested over speech detected by means of the Oracle VAD, hence all available speech in the *Test* subset. This comparison aims to test the SLOC accuracy in a absolute sense, independently from a VAD algorithm. As result, the data-driven model is more robust against the multi-room environment, outperforming the classical localization algorithm of more than 35% P_{cor}.

Oracle VAD		Average
Δ	RMS (mm)	-663
	P_{cor} (%)	+35.8

Table 7: Difference of the most performing SLOC proposed by the authors ($SLOC_{MC}{}^{\dagger}$) with the $SLOC_B$ in the presence of an Oracle VAD.

Finally, in Table 8 the overall performance of the proposed model and the baseline framework is reported in terms of difference. The comparison is reported in terms of Δ defined above. In terms of detection, a reduction of 3.0% SAD, of 4.2% FA and of 0.7% DeL is observed when the Joint-V $VAD^{\dagger}$ is employed. On the other hand, when the $SLOC_{MC}{}^{\dagger}$ is tested over true positive detected by the Joint-V $VAD^{\dagger}$, a higher accuracy on Pcor of 34.9% and a reduction on RMS of 589 mm is observed with respect to the $SLOC_B$.

		Average
Δ	SAD (%)	-3.0
	DeL (%)	-0.7
	FA (%)	-4.2
	RMS (mm)	-589
	P_{cor} (%)	+34.9

Table 8: Differences between the proposed data-driven approach and the baseline model of (Tachioka, Narita, Watanabe & Le Roux, 2014).

6. Conclusions

This work proposes a novel data-driven framework for detecting and localizing a speaker in a multi-room environment. For many years these two task have been studied as two separated problems, however their mutual dependency must be addressed in a real world scenario. This issue is dealt with within this work, where an architecture consisting of SLOC cascaded to VAD is proposed, and the dependency of the SLOC to VAD errors is investigated.

This work represents an extension of the authors' previous contribution (Vecchiotti, Principi, Squartini & Piazza, 2018), by introducing novel neural architectures for VAD and SLOC based on CNNs. In addition, the proposed framework is tested against the only other framework present in literature for the detection and localization of a speaker, which relies on classical VAD and SLOC algorithms and tackles the multi-room environment. The objective of the authors is to highlight the efficiency of a data-driven solution compared to a classical approach. Furthermore, the multi-room environment is here taken into account due to its high fidelity to real world applications.

In details, four CNN-based VAD algorithms are here compared, where the most performing one is capable of virtuously processing audio features commonly employed for VAD and SLOC, respectively. Two different SLOC architectures are then proposed, with the purpose of properly exploiting data recorded by multiple microphone installations.

In addition, the authors increased the quantity of training material by applying data augmentation. In particular, two subsets extend the original DIRHA dataset: the first one is another version of the employed dataset, while the second one is the result of an ad-hoc technique developed by the authors. In details, the RIRs of two virtual rooms equivalent in dimension to the rooms under observation have been generated, and speech data is then convolved with them. As result, when the proposed Joint-V VAD model has been trained with data augmentation technique, a SAD reduction of 3.0% is observed compared to the baseline work. Similarly, the data-driven SLOC architecture here discussed outperforms the reference framework in localization of a P_{cor} 34.9% higher and with a RMS 589 mm lower. The effectiveness of data augmentation is clearly observed for VAD and SLOC.

Future works will target the employment of new features for VAD and SLOC, especially aiming to a joint model performing simultaneous detection and local-

ization. Furthermore, it is in the interest of the authors to employ neural network characterized by a recurrent behaviour, also transfer learning techniques to adapt the models developed for certain rooms to other rooms, even related to different residential environments.

References

Allen, J. B., & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, *65*, 943–950.

Belloch, J. A., Gonzalez, A., Vidal, A. M., & Cobos, M. (2015). On the performance of multi-gpu-based expert systems for acoustic localization involving massive microphone arrays. *Expert Systems with Applications*, *42*, 5607 – 5620.

Benyassine, A., Shlomot, E., Su, H. ., Massaloux, D., Lamblin, C., & Petit, J. . (1997). ITU-T Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications. *IEEE Communications Magazine*, *35*, 64–73.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*, 41–75.

Chakrabarty, S., & Habets, E. A. P. (2017). Broadband DOA estimation using convolutional neural networks trained with noise signals. In *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 136–140).

Chollet, F. et al. (2015). Keras. `https://keras.io`.

Cristoforetti, L., Ravanelli, M., Omologo, M., Sosi, A., Abad, A., Hagmüller, M., & Maragos, P. (2014). The DIRHA simulated corpus. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)* (pp. 2629–2634).

Cui, X., Goel, V., & Kingsbury, B. (2015). Data augmentation for deep convolutional neural network acoustic modeling. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4545–4549).

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*, 357–366.

Do, H., Silverman, H. F., & Yu, Y. (2007). A real-time SRP-PHAT source location implementation using stochastic region contraction(SRC) on a large-aperture microphone array. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. I–121–I–124).

Ferguson, E. L., Williams, S. B., & Jin, C. T. (2018). Sound source localization in a multipath environment using convolutional neural networks. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2386–2390).

Ferroni, G., Bonfigli, R., Principi, E., Squartini, S., & Piazza, F. (2015). A deep neural network approach for voice activity detection in multi-room domestic scenarios. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

Fujimoto, M., & Ishizuka, K. (2008). Noise robust voice activity detection based on switching kalman filter. *IEICE Transactions on Information and Systems*, *91*, 467–477.

Hayashida, K., Morise, M., & Nishiura, T. (2010). Near field sound source localization based on cross-power spectrum phase analysis with multiple microphones. In *Proceedings of Interspeech* (pp. 2758–2761).

He, W., Motlicek, P., & Odobez, J.-M. (2018). Deep neural networks for multiple speaker detection and localization. In *Proceedings of International Conference on Robotics and Automation (ICRA)* (pp. 74–79).

Hughes, T., & Mierle, K. (2013). Recurrent neural networks for voice activity detection. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7378–7382).

Knapp, C., & Carter, G. (1976). The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*, 320–327.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5220–5224).

Kovandžić, M., Nikolić, V., Al-Noori, A., Ćirić, I., & Simonović, M. (2017). Near field acoustic localization under unfavorable conditions using feedforward neural network for processing time difference of arrival. *Expert Systems with Applications*, *71*, 138 – 146.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems (NIPS)* (pp. 1097–1105).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*, 2278–2324.

Lee, A., Nakamura, K., Nisimura, R., Saruwatari, H., & Shikano, K. (2004). Noise robust real world spoken dialogue system using GMM based rejection of unintended inputs. In *Proceedings of 8th International Conference on Spoken Language Processing (ICSLP)* (pp. 173–176).

Ma, N., May, T., & Brown, G. J. (2017). Exploiting deep neural networks and head movements for robust binaural localization of multiple sources in reverberant environments. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*, 2444–2453.

May, T., van de Par, S., & Kohlrausch, A. (2012). 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*, *20*, 2016–2030.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210).

Prisyach, T., Mendelev, V., & Ubskiy, D. (2016). Data augmentation for training of noise robust acoustic models. In *Proceedings of International Conference on Analysis of Images, Social Networks and Texts (AIST)* (pp. 17–25).

Ragni, A., Knill, K. M., Rath, S. P., & Gales, M. J. (2014). Data augmentation for low resource languages. In *Proceedings of Interspeech* (pp. 810 – 814).

Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, *24*, 279–283.

Schlüter, J., & Grill, T. (2015). Exploring data augmentation for improved singing voice detection with neural networks. In *Transactions of the International Society for Music Information Retrieval (ISMIR)* (pp. 121–126).

Seewald, L. A., Gonzaga Jr, L., Veronez, M. R., Minotto, V. P., & Jung, C. R. (2014). Combining SRP-PHAT and two kinects for 3D sound source localization. *Expert Systems with Applications*, *41*, 7106–7113.

Silva, D. A., Stuchi, J. A., Violato, R. P. V., & Cuozzo, L. G. D. (2017). Exploring convolutional neural networks for voice activity detection. In *Cognitive Technologies* (pp. 37–47). Springer International Publishing.

Sivasankaran, S., Vincent, E., & Campbell, D. R. (2017). Room impulse response generator. `https://github.com/sunits/rir_simulator_python`.

Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Processing Letters*, *6*, 1–3.

Tachioka, Y., Narita, T., Watanabe, S., & Le Roux, J. (2014). Ensemble integration of calibrated speaker localization and statistical speech detection in domestic environments. In *Proceedings of Hands-free Speech Communication and Microphone Arrays (HSCMA)* (pp. 162–166).

Taghizadeh, M. J., Garner, P. N., Bourlard, H., Abutalebi, H. R., & Asaei, A. (2011). An integrated framework for multi-channel multi-source localization and voice activity detection. In *Proceedings of Hands-free Speech Communication and Microphone Arrays (HSCMA)* (pp. 92–97).

Tashev, I., & Mirsamadi, S. (2016). Dnn-based causal voice activity detector. In *Information Theory and Applications Workshop*.

Tran, H. D., Ng, W. Z. T., & Leng, Y. R. (2017). Data augmentation, missing feature mask and kernel classification for through-the-wall acoustic surveillance. In *Proceedings of Interspeech* (pp. 3807–3811).

Tüske, Z., Golik, P., Nolden, D., Schlüter, R., & Ney, H. (2014). Data augmentation, feature combination, and multilingual neural networks to improve asr and kws performance for low-resource languages. In *Proceedings of Interspeech* (pp. 1420–1424).

Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., & Sarti, A. (2007). Scream and gunshot detection and localization for audio-surveillance systems. In *Proceedings of Conference on Advanced Video and Signal Based Surveillance (AVSS)* (pp. 21–26).

Vecchiotti, P., Principi, E., Squartini, S., & Piazza, F. (2018). Deep neural networks for joint voice activity detection and speaker localization. In *Proceedings of 26th European Signal Processing Conference (EUSIPCO)* (pp. 1567–1571).

Vesperini, F., Vecchiotti, P., Principi, E., Squartini, S., & Piazza, F. (2018). Localizing speakers in multiple rooms by using deep neural networks. *Computer Speech & Language*, *49*, 83–106.

Yantorno, R. E., Krishnamachari, K. R., Lovekin, J. M., Benincasa, D. S., & Wenndt, S. J. (2001). The Spectral Autocorrelation Peak Valley Ratio (SAPVR) - Usable Speech Measure Employed as a Co-channel Detection System. In *Proceedings of International Workshop on Intelligent Signal Processing (WISP)*.

Zhou, Y., Xiong, C., & Socher, R. (2017). Improved regularization techniques for end-to-end speech recognition. *arXiv preprint arXiv:1712.07108*.

Zöhrer, M., & Pernkopf, F. (2017). Virtual adversarial training and data augmentation for acoustic event detection with gated recurrent neural networks. In *Proceedings of Interspeech* (pp. 493–497).