Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment

Sunit Sivasankaran, Emmanuel Vincent, Dominique Fohr

Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment

Sunit Sivasankaran, Emmanuel Vincent, Dominique Fohr

Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

sunit.sivasankaran@inria.fr, emmanuel.vincent@inria.fr, dominique.fohr@loria.fr

Abstract

Speaker localization is a hard task, especially in adverse environmental conditions involving reverberation and noise. In this work we introduce the new task of localizing the speaker who uttered a given keyword, e.g., the wake-up word of a distant-microphone voice command system, in the presence of overlapping speech. We employ a convolutional neural network based localization system and investigate multiple identifiers as additional inputs to the system in order to characterize this speaker. We conduct experiments using ground truth identifiers which are obtained assuming the availability of clean speech and also in realistic conditions where the identifiers are computed from the corrupted speech. We find that the identifier consisting of the ground truth time-frequency mask corresponding to the target speaker provides the best localization performance and we propose methods to estimate such a mask in adverse reverberant and noisy conditions using the considered keyword.

Index Terms: Speaker localization, wake-up word, convolutional neural network, reverberation, overlapping speech.

1. Introduction

Speaker localization is the task of estimating the direction of arrival (DOA) of speech uttered by a speaker [1]. This is useful for various applications such as speech enhancement and separation [2–7] or robotic sensing [8]. DOA estimation is usually done in the short time Fourier transform (STFT) domain [9]. For two-channel data, a simple approach is to compute the time difference of arrival (TDOA) in each time-frequency bin and to find the peak of the resulting TDOA histogram [10]. Techniques such as generalized cross-correlation with phase transform (GCC-PHAT) [11] and multiple signal classification (MUSIC) [12] estimate the DOA by finding the peak of a so-called angular spectrum instead [13–16]. These techniques assume that each time-frequency bin is dominated by the direct component of a single source. Their performance degrades in adverse environmental conditions involving reverberation or noise.

To improve the robustness of DOA estimation, deep neural networks (DNNs) have been proposed to learn a mapping between signal features and a discretized DOA space [17–21]. Various features such as phasemaps [17, 18] and GCC-PHAT [21] have been used as inputs. In [22], the cosines and sines of the frequency-wise phase differences between microphones, termed as cosine-sine interchannel phase difference (CSIPD) features, have been shown to perform as well as phasemaps for DOA estimation, despite their lower dimensionality. We hence use the latter features in this work.

In the presence of multiple speakers, localization becomes harder due to the nonlinear nature of phase mixing. Nevertheless, due to the approximate disjointness of speech signals in the time-frequency plane [23], most time-frequency bins are dominated by a single speaker. This property has motivated clustering-based localization algorithms, which iteratively identify the time-frequency bins dominated by each speaker and reestimate the corresponding DOAs [24–26]. It has also recently been exploited to design training data for multi-speaker DNN-based localization [18, 19].

In the following, we propose to identify the DOA of a single speaker in a multi-speaker distant-microphone voice command scenario by exploiting knowledge of the wake-up word. The wake-up word can be identical for all speakers, but only a single speaker (henceforth referred to as the *target*) can utter it at a given time. The wake-up word is phonetically aligned with the speech signal using an automatic speech recognition (ASR) system. The phonetic alignments are then used to obtain information (henceforth referred to as *target identifier*) about the target speaker which is used as input together with CSIPD features in order to estimate the target DOA. This work contrasts with earlier works on multi-speaker localization which aimed to estimate the DOAs of all speakers. Identifying the target would then require additional post-processing which can be error-prone. Though identifying the direction of all the speakers has its own utility, identifying the DOA of the target is useful in applications such as speech recognition in the presence of overlapping speech, where the recognition can be restricted to speech produced by the target. To the best of our knowledge, this problem has not been studied in the literature before.

In the following, we assess the utility of various spectrum-based or mask-based target identifiers. Spectrum-based identifiers include the clean, early reverberated, and fully reverberated magnitude spectrum of the target speech, and the "phoneme spectrum" [27] which is the average clean spectrum corresponding to the spoken phoneme. Mask-based identifiers measure the proportion of sound magnitude attributed to the target signal in each time-frequency bin. We evaluate the localization performance in real scenarios where the identifiers are estimated from the corrupted speech given the spoken keyword.

Notations are introduced in Section 2 and the proposed target identifiers in Section 3. The experimental setup is described in Section 4 and the results are reported in Section 5. We conclude in Section 6.

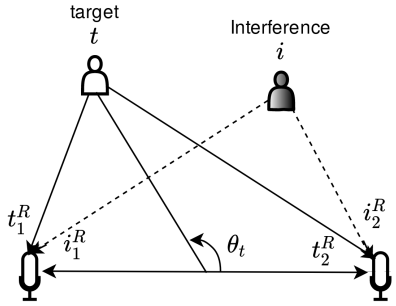
2. Problem description

The problem setup is shown in Fig. 1. Two microphones are placed inside a room. A target and an interfering speaker are placed at azimuth angles θ_t and θ_i with respect to the microphone axis, respectively. Denoting by t the target speech signal and by i the interfering speech signal, the signal received at the c-th microphone is:

$$s_c = t_c^R + i_c^R + \eta_c \tag{1}$$

where $t_c^R = r_c^t \star t$ and $i_c^R = r_c^i \star i$ are the reverberated target and interfering signals received at the c-th microphone. Here,

Figure 1: *Problem setup.*



r_c^t and r_c^i are the room impulse responses (RIR) of the target and the interference at the c^{th} microphone, η_c is the noise, and $\star$ denotes the convolution operator. The goal is to estimate θ_t using s_c and the keyword spoken by the target.

To achieve this goal, we use a convolutional neural network (CNN) to estimate θ_t given features extracted from s_c as inputs. Specifically, we use the CSIPD features [22]

$$\text{CSIPD}[\omega, n] = [\cos(\Delta\phi[\omega, n])\ ,\ \sin(\Delta\phi[\omega, n])] \quad (2)$$

for all time frames n and frequency bins ω, where

$$\Delta\phi[\omega, n] = \angle S_1[\omega, n] - \angle S_2[\omega, n]. \quad (3)$$

Here S_1 and S_2 are the STFT coefficients of the signals received at the microphones and $\angle$ denotes the phase of a complex number. The size of the input feature vector in each time frame is twice the number of STFT bins. Similar to [22], we quantize the output DOA space into 181 classes.

3. Target identifiers

In this section we describe a set of target identifiers which can be used as additional inputs to the CNN in order to help it focus on the target. These identifiers can broadly be categorized into spectrum-based or mask-based identifiers. Spectrum-based identifiers are estimates of the magnitude spectrum of the target signal, while mask-based identifiers are soft time-frequency masks. We explain the extraction of these identifiers in an ideal situation where the target signal is known and in real scenarios where the identifiers are extracted from the corrupted speech s_c.

3.1. Spectrum-based identifiers

The magnitude spectrum $|T|$ of the uncorrupted target signal t is not corrupted by either interference or noise. This makes it an ideal target identifier which we refer to as the *clean spectrum identifier*. Extracting this identifier from the mixture s_c is hard, however. The reverberated spectrum $|T^R|$ is easier to estimate in practice. We refer to this spectrum as the *reverberated spectrum identifier*. The reverberated speech t_c^R is corrupted by late reverberation which is known to have a detrimental effect on single-speaker DOA estimation [17, 22]. Therefore, we also consider the spectrum $|T^E|$ of the signal containing only the direct component and the early reflections of the target signal as an identifier and call it the *early spectrum identifier*. The latter signal is computed by convolving speech with the the first τ_E samples of the RIRs:

$$t_c^E = r_c^t[0 : \tau_E] \star t. \quad (4)$$

The spectra are averaged over all microphones. For instance,

$$|T^E| = \frac{1}{C}\sum_c |T_c^E| \quad (5)$$

with $|T_c^E|$ the magnitude spectrum of t_c^E.

Another quantity of interest, which relies on the keyword spoken by the target speaker rather than the target signal itself, is the average clean spectrum corresponding to that keyword. This quantity was introduced in [27] in the context of speech enhancement. To obtain it, a first ASR system is trained on a large clean speech corpus and used to obtain phonetic alignments for the training data and to compute the average spectrum for every phonetic class in that data. A second ASR system is trained (possibly on a distinct, reverberated or noisy corpus) and used to obtain a phonetic alignment for the mixture signal s_c for the known keyword. The average spectrum corresponding to the phonetic class in each time frame is then retrieved and called the *phoneme spectrum identifier*.

3.2. Mask-based identifiers

As an alternative to the above spectrum-based identifiers, mask-based identifiers represent the ratio of the magnitude spectrum of the target divided by the sum of the magnitude spectrum of the target and that of other sounds. This ensures that the resulting mask always lies between 0 and 1. For instance, the *early mask identifier* is computed as follows:

$$\delta_c = s_c - t_c^E \quad (6)$$

$$|\Delta| = \frac{1}{C}\sum_c |\Delta_c| \quad (7)$$

$$M^E = \frac{|T^E|}{|T^E| + |\Delta|} \quad (8)$$

where $|\Delta_c|$ is the magnitude STFT of δ_c. The *clean mask identifier* and *reverberated mask identifier* can be obtained in a similar way by replacing the early target spectrum ($|T^E|$) with the clean spectrum ($|T|$) or the reverberated spectrum ($|T^R|$), respectively.

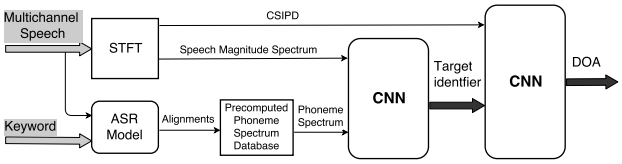
3.3. Appending vs. multiplication

The target identifiers are appended to the CSIPD features and used as inputs to the CNN both for training and testing. The network will learn to correlate the target identifier time-frequency bins with the CSIPD features while estimating the DOA. In the particular case of the target mask identifiers, this relationship can be directly imposed on the features by multiplying the CSIPD features with the mask in every time-frequency bin, thereby freeing the network from explicitly learning such a relationship. This would not be possible in the case of phasemap features.

3.4. Estimating the identifiers using the target keyword

A schematic representation of the target identifier estimation process in realistic scenarios, without access to the target clean speech, is shown in Fig. 2. The ASR system uses the speech signal along with the known keyword to compute the phoneme level alignment. The sequence of phoneme spectra corresponding to this alignment is retrieved from a precomputed list of phoneme spectra. These spectra are concatenated with the magnitude spectra of all channels of the corrupted speech signal in order to form a sequence of feature vectors. The shape of each

Figure 2: *Block diagram describing the computation of target identifiers and DOA estimation.*



feature vector is $(C+1) \times D$ with D the number of frequency bins. A CNN then learns a mapping between the input features and the target identifier in each time frame. All spectrum-based identifiers (except for the phoneme spectrum identifier which does not need any further estimation) and mask-based identifiers are obtained using this technique.

The CNN used to estimate the identifier consists of four convolutional layers followed by a fully connected layer of dimension 512, leading to an output layer of dimension D. The first, second, third and fourth convolutional layers contain 64, 32, 16, and 8 feature maps, respectively, each with filter shape 3×1. Max pooling of shape 2×1, batch normalization [28] and dropout [29] are used in all convolutional layers. Note that the CNN operates on a single frame (no convolution over the time axis). Rectified linear unit (ReLU) nonlinearities are used in all hidden layers. For spectrum-based identifiers, a linear layer is used at the output. For mask-based identifiers, a sigmoid nonlinearity is used at the output instead. All networks were trained using Adam [30] using the mean squared error (MSE) as the cost function. The network is trained for 100 epochs with a minibatch size of 512. The model corresponding to the epoch with the smallest MSE score in the development set is retained.

4. Experimental setup

In this section we detail our experimental setup.

4.1. Creating RIRs

All experiments in this paper were conducted using two microphones ($C = 2$). A shoebox model based on the image source method was used to simulate RIRs via RIR-Generator [31]. For every configuration, a room with random dimensions varying from 3 m to 9 m was chosen. The reverberation time was picked randomly in the range of $[0.3, 1]$ s. Two microphones were placed inside the room at a distance of 10 cm. This was done by randomly positioning the first microphone inside the room and then positioning the second microphone by selecting a random point on a sphere of radius 10 cm with the first microphone as the center. The positions of the microphones were ensured to be at least 50 cm from any wall. In order to position the target or the interference at a given angle, a point was randomly chosen on the surface of the circular cone whose axis is the microphone axis, whose center is the midpoint of the microphone line, and whose angle is the desired target or interference DOA. The DOA space was quantized into 1° classes. The possible DOA range is therefore $[0, 180]^\circ$, implying 181 classes. A minimum separation of 5° between the target and the interference was ensured. For every possible pair of target and interference DOAs, 50, 1, and 2 such room configurations were generated for training, development, and test, respectively. This resulted in 1,557,600, 31,152, and 62,304 different configurations for training, development, and test, respectively. For every configuration, the distance between the target or interfering speaker and the microphones was randomly chosen in the range of $[0.5, 5.5]$ m. In many cases, the interfering speaker is closer to the microphones than the target. The resulting frame-wise direct-to-reverberant ratio (DRR) values are in the range of $[-40, +12]$ dB for the test set with an average of -6 dB. This can be considered as a challenging, realistic scenario.

4.2. Signal generation and feature extraction

Speech utterances for simulating target and interference speech were picked from the Librispeech [32] dataset. They were divided into training, development, and test sets with no overlap. Two different speech signals were convolved with the simulated RIRs of a single room to obtain the target and interference speech components of the mixture s_c. These components were combined at a random signal-to-interference-ratio (SIR) in the range of $[0, +10]$ dB. Speech-shaped noise (SSN) [33–35] was added to this mixture at a random signal-to-noise-ratio (SNR) in the range of $[0, +15]$ dB for the training set and $[0, +30]$ dB for the development set. The SSN signals were created by filtering white noise with a filter whose frequency response was computed by averaging the magnitude spectra of 3,000 STFT frames (different for every mixture). Excerpts of real ambient noise from the voiceHome corpus [36] were included in the test set instead of SSN. These stationary diffuse noises were recorded using a microphone pair with 10 cm spacing in three different apartments in a similar fashion as in [37]. Nonoverlapping time frames with 100 ms duration containing unique but nonidentical phonemes for target and interference were extracted to compute the features. Phoneme-level alignments obtained from the underlying clean speech signal were used to determine the presence and duration of a particular phoneme. The alignments were delayed to account for the delay due to sound propagation from the target speaker to the microphone pair. For every time frame, a sine window was applied and a 1,600 point Fourier transform was computed. The resulting CSIPD features had an overall dimension of $2D = 1{,}602$. Only a single time frame was kept for every training and development utterance in order to maximize the diversity of the training and development sets. For every test utterance, a sequence of $N = 15$ time frames was kept instead: this corresponds to a signal duration of 1.5 s which is a typical duration for a wake-up word.

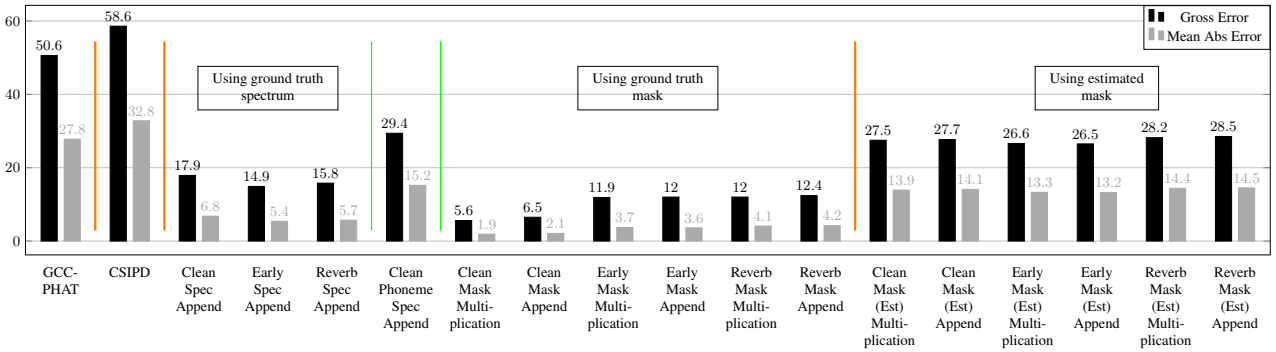
To obtain the clean spectrum identifier, the clean target signal t was scaled by the direct component of the RIR. The early spectrum identifier was obtained by convolving the target signal with the first $\tau_E = 50$ ms of the RIR. Finally, the reverberated spectrum identifier was obtained by convolving the target signal with the non-truncated RIR. The phoneme spectra were computed from the clean Librispeech training set. The mask-based identifiers were computed using the same spectra. All spectra are of dimension $D = 801$.

4.3. ASR system, network architecture, and DOA pooling

Phoneme-level alignments were obtained via an HMM-GMM based system trained on clean speech using speaker-adapted Mel-frequency cepstral coefficients (MFCC) [38]. This system was applied to the clean training, development, and test data in all our experiments.

The CNN architecture used to estimate the DOA is similar to the one used to estimate the target identifier (see Section 3.4). Four convolutional layers of size 64, 32, 16, and 8 followed by a fully connected hidden layer with ReLU units are used. They are connected to a softmax output layer with 181 classes. Max

Figure 3: *Gross error rate (%) and mean absolute error (°) of DOA estimation. The bar chart is divided into four zones separated by orange lines, which show respectively the results obtained by GCC-PHAT, CNN using CSPID features alone, CNN using CSIPD features with various ground truth target identifiers, and CNN using CSIPD features with various estimated target identifiers.*



pooling, batch normalization, and dropout are used in all convolutional layers. Cross-entropy is used along with the Adam optimizer [30] to train the network. The input dimension is either $3 \times 801 \times 1$ when the target identifier is appended to the CSIPD features or $2 \times 801 \times 1$ when the identifier is multiplied by the CSIPD features or there is no identifier.

The CNN operates on a single frame (no convolution over time). It outputs a posterior DOA distribution denoted as $p(\theta[n])$ in each time frame n. The target DOA for the whole utterance is then found as $\widehat{\theta} = \arg\max_\theta \max_{n \in \{1,\dots,N\}} p(\theta[n])$. This form of max pooling worked better than averaging individual DOA estimates across all frames.

5. Results and discussion

Two different metrics are used to evaluate the estimation performance on the test set, namely, the gross error rate which measures the percentage of utterances whose estimated DOA is above an error threshold (set to $5°$) and the mean absolute error which is the average absolute DOA estimation error in degrees over all utterances.

The results obtained using the proposed CNN and its comparison with GCC-PHAT are shown in Fig. 3. GCC-PHAT results in a gross error rate of 50.6% and a mean absolute error of $27.8°$. The CNN applied to CSIPD features with no target identifier yields a gross error rate of 58.6% which is worse than GCC-PHAT. As expected, this shows that localizing a single-speaker in a multi-speaker scenario without exploiting specific information about the target speaker is infeasible.

Using the clean, early, and reverberated spectrum identifiers as additional inputs to the CNN reduces the gross error rates down to 17.9%, 14.9%, and 15.8%, respectively. This indicates that these identifiers encode target information. The phoneme spectrum identifier results in a gross error rate of 29.4%, that is a 99% relative improvement over using CSIPD features alone. This shows that the phoneme information obtained by aligning the keyword can be directly used to identify the target.

The ground truth mask-based identifiers result in the best localization performance. The lowest gross error rates of 5.6% and 6.5% are observed by multiplying and appending the clean mask with CSIPD features, respectively. The low mean absolute error of $1.9°$ shows the relevance of the clean mask as a target identifier. Slightly larger gross error rates of 11.9% and 12.0% are obtained by multiplying the CSIPD features with the early and reverberated masks, respectively. Multiplying the mask generally gives better performance than appending the mask. This may be because the identifier information is directly encoded in the input features while the network is forced to learn how to exploit this additional information when the mask is appended.

Due to the fact that mask-based identifiers outperform spectrum-based identifiers in the ground truth setting, we only estimated mask-based identifiers in the real setting using the approach in Section 3.4. The best gross error rate of 26.5% is obtained by estimating the early mask target identifier from the corrupted speech and appending it with the CSIPD features. This is better than the performance obtained using the phoneme spectrum identifier, which indicates that additional information was learned during the mask estimation process.

6. Conclusions

In this work, we proposed a method to localize a single speaker in a multi-speaker environment using additional identifier information. We investigated multiple such identifiers based on speech spectra or masks and explained how to estimate them using the phonetic information extracted from the keyword spoken by the target speaker. The best localization performance in a realistic scenario was obtained by concatenating CSIPD features with a mask representing the proportion of direct sound and early reflections from the target speaker. Although this resulted in improved localization performance, the performance remains significantly lower when compared to the ground truth mask. Future work will deal with methods aiming to estimate a better mask in an end-to-end fashion with the goal of improving the localization performance. We will also consider full-fledged CNNs involving convolution over time and assess the impact of estimating phoneme-level alignments from corrupted speech rather than clean speech.

7. Acknowledgements

This work was made with the support of the French National Research Agency, in the framework of the project VOCADOM "Robust voice command adapted to the user and to the context for AAL" (ANR-16-CE33-0006). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see `https://www.grid5000.fr`).

8. References

[1] M. S. Brandstein and D. B. Ward, Eds., *Microphone Arrays: Signal Processing Techniques and Applications*. Springer, 2001.

[2] M. Wölfel and J. McDonough, *Distant Speech Recognition*. Wiley, 2009.

[3] I. Cohen, J. Benesty, and S. Gannot, Eds., *Speech Processing in Modern Communication: Challenges and Perspectives*. Springer, 2010.

[4] T. Virtanen, R. Singh, and B. Raj, Eds., *Techniques for Noise Robustness in Automatic Speech Recognition*. Wiley, 2012.

[5] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, *Robust Automatic Speech Recognition*. Academic Press, 2015.

[6] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[7] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, 2018.

[8] C. Rascon and I. Meza, "Localization of sound sources in robotics: A review," *Robotics and Autonomous Systems*, vol. 96, pp. 184–210, Oct. 2017.

[9] Ö. Yılmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.

[10] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *The Journal of the Acoustical Society of America*, vol. 116, no. 5, pp. 3075–3089, Nov. 2004.

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

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

[13] F. Nesta, P. Svaizer, and M. Omologo, "Cumulative state coherence transform for a robust two-channel multiple source localization," in *Independent Component Analysis and Signal Separation*, Mar. 2009, pp. 290–297.

[14] N. Roman, D. Wang, and G. J. Brown, "Speech segregation based on sound localization," *The Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, Oct. 2003.

[15] Z. El Chami, A. Guérin, A. Pham, and C. Servière, "A phase-based dual microphone method to count and locate audio sources in reverberant rooms," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2009, pp. 209–212.

[16] C. Liu, B. C. Wheeler, W. D. OBrien Jr, R. C. Bilger, C. R. Lansing, and A. S. Feng, "Localization of multiple sound sources with two microphones," *The Journal of the Acoustical Society of America*, vol. 108, no. 4, pp. 1888–1905, 2000.

[17] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2017, pp. 136–140.

[18] ——, "Multi-speaker localization using convolutional neural network trained with noise," in *NIPS 2017 Workshop on Machine Learning for Audio Processing*, Dec. 2017.

[19] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *2016 IEEE Spoken Language Technology Workshop*, Dec. 2016, pp. 603–609.

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

[21] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *2016 IEEE International Workshop on Machine Learning for Signal Processing*, Sep. 2016, pp. 1–6.

[22] V. Varanasi, R. Serizel, and E. Vincent, "DNN based robust DOA estimation in reverberant, noisy and multi-source environment," in preparation.

[23] S. Rickard and Ö. Yılmaz, "On the approximate W-disjoint orthogonality of speech," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2002, pp. I–529–I–532.

[24] M. I. Mandel, D. P. W. Ellis, and T. Jebara, "An EM algorithm for localizing multiple sound sources in reverberant environments," in *19th International Conference on Neural Information Processing Systems*, 2006, pp. 953–960.

[25] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1592–1604, Jul. 2007.

[26] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, pp. 1950–1960, Mar. 2012.

[27] Z. Chen, S. Watanabe, H. Erdogan, and J. R. Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Interspeech*, 2015.

[28] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd International Conference on Machine Learning*, vol. 37, Jul. 2015, pp. 448–456.

[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2014.

[31] E. A. P. Habets, "RIR-Generator: Room impulse response generator." [Online]. Available: https://github.com/ehabets/RIR-Generator

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr. 2015, pp. 5206–5210.

[33] M. Pariente and D. Pressnitzer, "Predictive denoising of speech in noise using deep neural networks," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 2611–2611, Oct. 2017.

[34] F. Li, P. S. Nidadavolu, and H. Hermansky, "A long, deep and wide artificial neural net for robust speech recognition in unknown noise," in *Interspeech*, 2014, pp. 358–362.

[35] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Speech enhancement for a noise-robust text-to-speech synthesis system using deep recurrent neural networks," in *Interspeech*, 2016, pp. 352–356.

[36] N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, É. Lamandé, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, S. Fleury, and E. Jamet, "A French corpus for distant-microphone speech processing in real homes," in *Interspeech*, 2016, pp. 2781–2785.

[37] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–19, 2016.

[38] M. J. F. Gales, "Maximum likelihood linear transformations for HMM based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.