# No Need to Scream: Robust Sound-based Speaker Localisation in Challenging Scenarios

Tze Ho Elden Tse[1], Daniele De Martini[2][0000−0001−6121−5839], and Letizia Marchegiani[3][000−0001−6782−6657]

[1] Electronic Systems, BAE Systems, UK eldentseb@gmail.com
[2] Oxford Robotics Institute, University of Oxford, UK daniele@robots.ox.ac.uk
[3] Department of Electronic Systems, Aalborg University, Aalborg Ø, DK lm@es.aau.dk

**Abstract.** This paper is about speaker verification and horizontal localisation in the presence of conspicuous noise. Specifically, we are interested in enabling a mobile robot to robustly and accurately spot the presence of a target speaker and estimate his/her position in challenging acoustic scenarios. While several solutions to both tasks have been proposed in the literature, little attention has been devoted to the development of systems able to function in harsh noisy conditions. To address these shortcomings, in this work we follow a purely data-driven approach based on deep learning architectures which, by not requiring any knowledge either on the nature of the masking noise or on the structure and acoustics of the operation environment, it is able to reliably act in previously unexplored acoustic scenes. Our experimental evaluation, relying on data collected in real environments with a robotic platform, demonstrates that our framework is able to achieve high performance both in the verification and localisation tasks, despite the presence of copious noise.

**Keywords:** Speaker Localisation · Speaker Verification · Speech in Noise.

## 1 Introduction

Human-robot social interaction greatly relies on accurate detection and localisation of the various interlocutors [28]. In many instances, this interaction requires voiced communication (*e.g.* a robot personal assistant executing commands uttered by a specific user). Such social contexts, however, can be characterised by a high levels of noise of a various nature, including the noise emitted by the robot while moving. For instance, elderly people trying to communicate with a robot might struggle to be perceived, because they might be simultaneously watching TV at a high volume, there might be other people talking, or they might have developed, through the years, non-coordination or weakness of the speech mechanism, resulting in the production of very feeble sounds. This paper explores the possibility of spotting and localising a specific speaker in challenging scenarios, characterised by a significantly low Signal-to-Noise Ratio (SNR) level, relying on

the use of Convolutional Neural Networks (CNNs). Specifically, we are interested in enabling a mobile robot assistant to spot the presence of its intended user in the acoustic scene (known in the literature as *speaker verification*), and estimate his/her position on the horizontal plane (*i.e. speaker localisation*).

Literature in robotics has provided us with several speaker detection and recognition frameworks, most of which rely on face and voice characteristics (see [19,26] among others). Despite the accuracy of such frameworks, situations where noise might play a crucial role and heavily compromise the quality of the sound perceived by the robot have not been yet investigated. This paper aims to move a step in this direction by exploring the performance of an audio-only perceptual system when challenged by extreme environmental conditions, which are not known *a priori*. As the conditions of the application scenarios are not necessarily either predictable or available, verification and localisation cannot take into account and leverage the structure of the operation environment, or the nature and characteristics of the overlapping noise. To address such constraints, we build a dataset containing several kinds of potential maskers, combined in a different fashion, and we opt for a pure data-driven approach, in the attempt of developing a framework able to generalise to unexplored acoustic scenes. We focus on binaural perception (*i.e.* stereo signals), and thus, on horizontal localisation (*i.e.* estimation of the direction of arrival of the sound on the horizontal plane), as a *proof-of-concept*. Yet, additional audio channels as well as sensing modalities could be considered to extend the analysis to 3D localisation. Despite the use of deep learning frameworks which, traditionally, require a large amount of training data, in our scenario we are able to obtain remarkable performance relying only on 30 min of *target speech*, allowing the use of such technology also in scenarios where data collection might be particularly challenging (*e.g.* nursing homes). Our experimental evaluation proves that our system is able to accurately spot the presence of a specific speaker in the acoustic scene with an average verification rate of 94 %, and a median localisation error lower than 6°.

## 2 Related Works

**Speaker Verification** While traditionally this task has been addressed relying on Gaussian Mixture Models (*e.g.* [19,21]), recent advances in machine learning, particularly in the form of deep learning architectures (*e.g.* [8,9]) have dictated and driven the development of new methods able to achieve great precision, and to overcome the need of defining hand-crafted features. Several speaker verification systems, for instance, both text-dependent and text-independent have been introduced (*cf.* [2,25] among others). While those frameworks show particularly high accuracy in the classification process, not much can be said on how well they operate in harsh noisy conditions. Our work, which aims to verify the presence of a target speaker in the acoustic scene without relying on the use of specific text (*i.e.* text-independent), shares the aspirations of [25] and [2], and, as [2] explores the possibility of treating the acoustic signals as images to which directly apply CNNs. With respect to [2], however, in this paper we further challenge the

verification procedure, by considering scenarios characterised by the presence of heavy noise, of a a different nature and coming from different sources. The goal is to investigate the robustness of similar systems when acting in the real-world, when the presence of maskers of variegate types can, indeed, be plenteous.

**Speaker Localisation** Speaker or, more generally, sound source localisation, has followed a similar pattern, and more traditional geometrical methods [19,22] have been now superseded by deep learning approaches, such as [7, 14]. Both those studies rely on cross-correlation information to train CNN-based models to perform localisation. We are close in spirit to both those instances; but, as in the case of the speaker verification task, we wish to explore the behaviour of our system when coping with remarkably noisy scenes (*e.g.* while in [14], $SNR \geq 0dB$, in this work we consider scenarios with $-5dB \leq SNR \leq -20dB$), where the competitive maskers can be of a different nature and not only represented by other speakers in known indoor scenes, to evaluate its performance beyond laboratory and restrained environments. Furthermore, rather than utilising cross-correlation information, we propose the use of a stereo spectral representation of the signals, based on the Gammatone filterbanks (*cf.* Section 3.1).

In summary, the use of deep learning for speaker verification and localisation has been already investigated in the literature. Yet, with respect to those studies, this paper offers three main contributions:

- we consider challenging scenarios where $-5dB \leq SNR \leq -20dB$, while previous works only operate on positive levels of SNR;
- we consider both indoor and outdoor scenarios, and propose solutions able to operate robustly in both situations, overcoming any dependence on the structure of the operation environment and on the noise's characteristics;
- we propose the use of spectral stereo features, rather than ones based on cross-correlation, generally used for speaker localisation, to better cope with the presence of massive noise.

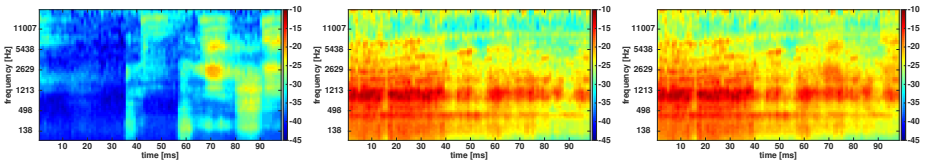
## 3 Technical Approach

Similarly to previous works in the area [19], our framework relies on a two-stage approach: firstly the incoming audio signal is fed to a CNN to verify the presence of the target speaker; secondly, in case the target speaker is present, further analysis is performed to estimate the Direction of Arrival (DoA) of the sound (*i.e.* horizontal localisation). A description of the feature representation of the various signals is given in Section 3.1, while the deep learning architectures employed are illustrated in Section 3.2.

### 3.1 Feature Representation

Traditionally, speaker recognition and verification tasks have been performed employing Mel-Frequency Cepstrum Coefficients (MFCCs) as feature representations of the audio signals; yet, recent works (*e.g.* [3, 18]) demonstrated that

the discriminative power of MFCCs greatly decreases when dealing with more realistic, dynamic and complex noisy scenarios. In this work, we adopt gammatonegrams, which are a visual representation of the energy of a signal based on short-time Fourier transform (STFT) and the application of Gammatone filterbanks [13], which have been firstly introduced in [11]. It has been proved, indeed, that such filtering, is able to guarantee robustness to noise for speech analysis tasks [15, 24]. The gammatonegrams are generated following the specification illustrated in [29] and [17], employing a bank of 64 filters. Examples of gammatonegrams from our dataset representing, respectively, target speech, noise and their combination at −20 dB are reported in Figure 1.




(a) Target Speech (b) Noise (c) Target Speech in Noise

Fig. 1: Example of the gammatonegram representation of sound frames of 1 s. From left to right: target speech, noise, and their combination at −20 dB. The energy of the time-frequency bins is expressed in decibel (dB).

### 3.2 Architecture

We approach both tasks using CNN-based models. We refer to the CNN used for verification as *VER-MONO*, and to the one used for localisation as *LOC-STEREO*. The two architectures differ in the input size and, consequently, in the size of the fully-connected layer, and in the structure of the output layer. Indeed, while the former takes the gammatonegram of a single channel as input, the latter considers a stereo gammatonegram, where the gammatonegrams corresponding the the audio channels are disposed *side by side*. Further details are provided in Figure 2. Furthermore, different loss functions are utilised in the training phase: while in the case of *VER-MONO*, we optimise a soft-max combined with a cross-entropy loss function to implement classification (*i.e.* target speaker vs anything/anyone else), in the case of the *LOC-STEREO* network, we minimise the Euclidean loss between the estimated DoA of the sound and the ground truth value.

## 4 Experimental Evaluation

We performed four kinds of experiments to validate our framework.

– Experiment 1: we train the *VER-MONO* network and evaluate its performance at verifying the presence of the target speaker against other random speakers, robot's, and environmental noise.

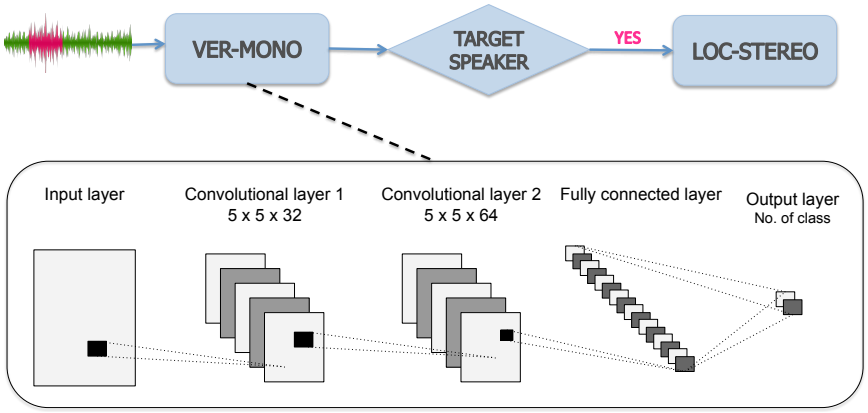



Fig. 2: The figure reports the two-stage approach operating in the framework: firstly the incoming audio signal is fed to a CNNs to verify the presence of the target speaker (*VER-MONO*) through a binary classification paradigm; secondly, in case the target speaker is present, further analysis is performed to estimate the DoA of the sound through a regression model (*LOC-STEREO*). The figure reports details on the architecture used for verification, which consists of two $5 \times 5$ convolutional layers, followed by a $2 \times 2$ max pooling, and one fully connected layer. All layers are equipped with a Rectified Linear Unit (ReLU). The regression employs the same architecture, but while *VER-MONO* operates on the gammatonagram of only one of the two channels available, *LOC-STEREO* acts on the gammatonegrams of both channels, disposed side by side. The fully-connected layer will have, thus, a different size as well. Lastly, the regression network counts only one unit in its output layer.

– Experiment 2: we compare the behaviour of the *VER-MONO* network with a different one, having the same architecture, but where verification is performed on a stereo gammatonegram (organised side by side, as in *LOC-STEREO*), and which we name *VER-STEREO*. The goal is to investigate whether having a stereo combination of the audio signals might help the network in the verification process.

– Experiment 3: we train the *LOC-STEREO* network, and evaluate its performance in the DoA estimation of the speaker's voice.

– Experiment 4: we compare the behaviour of the *LOC-STEREO* network with a different one, having the same architecture, but where localisation is performed on the cross-correlation between the gammatonegrams corresponding to the audio signals in the two channels, named *LOC-CROSS* and used as a benchmark. We also compare the behaviour of *LOC-STEREO* and *LOC-CROSS* with a more traditional geometrical method [19], indicated as *BASELINE*.

### 4.1 The Dataset

To evaluate our framework, we built a dataset where the voices of three speakers are combined with several kinds of noise, at different level of SNRs. Specifically, we consider particularly challenging scenarios characterised by significantly low SNRs: $SNR \in \{-5dB, -15dB, -20dB\}$. As our goal is enabling mobile robots to spot the presence of a speaker of reference in the acoustic scene, and localise him/her on the horizontal plane, the speech utterances were recorded with the speaker standing at different angles with respect to the front of the robot, and the sound emitted by the motors of the robot was also considered as one of the potential sources of noise (*i.e.* the robot might move to follow the speaker, or being already moving while it encounters the voice of the speaker). The data was recorded by using two Knowles omnidirectional boom microphones, mounted in proximity of each of the two front wheels of a *Clearpath Husky A200* platform (shown in Figure 3), and an ALESIS IO4 audio interface, at a sampling frequency of 44.1 kHz and a resolution of 16 bits. The speakers' voices were recorded in a silent, but realistic environment (*i.e.* perturbations due to reflections etc are present), accounting for a total of around 90 min minutes of speech data (*i.e.* 30 min minutes per speaker used as a reference). In addition to the robot's noise, we also considered other sources of stereo environmental noise from the *Urban Sound Dataset* [23], and from other publicly available databases, such as `www.freesound.org`, as well as random speech from [6]. Care was taken so that the environmental noise selected is characterised by moving sound sources, and covering both outdoor and indoor scenarios. Following previous works (*e.g.* [4] [16] [20]), we opted for collecting the data corresponding to the target speech and the masking noise separately, because this would allow us to directly control the level of SNR, and to isolate and accurately quantify the impact of the noise on the verification and localisation tasks.

### 4.2 Implementation Details

We trained the networks using mini-batch gradient descent based on back-propagation, employing the Adam optimisation algorithm [12]. Dropout [10] was applied to each layer for both architectures with a keeping probability of 0.75. The models were implemented using the Tensorflow [1] libraries. Similarly to previous works on deep learning in the auditory domain (cf. [5], [27]), we randomly split our dataset into training set (70%) and testing set (30%).

### 4.3 Experiment 1 and 2: Speaker Verification

We perform speaker verification at different SNR levels ($SNR \in \{-5dB, -15dB, -20dB\}$), comparing the behaviour of the *VER-MONO* and *VER-STEREO* networks, when operating on frames of $1s$. In both cases, the verification is implemented as a binary classification problem, where one class refers to a combination of the target speaker and different kinds of noise (as described in Section 4.1), and the other class is obtained as random combinations of urban noise, robots'

noise and voices from one or more speakers different from the target one. Table 1 reports the results of those experiments. We observe that both networks are characterised by high accuracy, with the *VER-MONO* providing slightly greater performance. This suggests that the verification process does not benefit from the use of stereo information.

Fig. 3: Clearpath Husky A200 used in the experiments.

| | Verification Accuracy | |
|---|---|---|
| SNR | *VER-MONO* | *VER-STEREO* |
| -5dB | **99.1 ± 0.80** | 99.88 ± 0.56 |
| -15dB | **98.65 ± 0.46** | 95.20 ± 0.82 |
| -20dB | **86.02 ± 2.42** | 82.26 ± 2.21 |
| Average | **94.85** | 92.19 |

Table 1: Verification accuracy (mean and variance), when employing the *VER-MONO* and the *VER-STEREO* networks in scenarios characterised by different SNRs. Accuracy is expressed as the percentage of correct verifications.

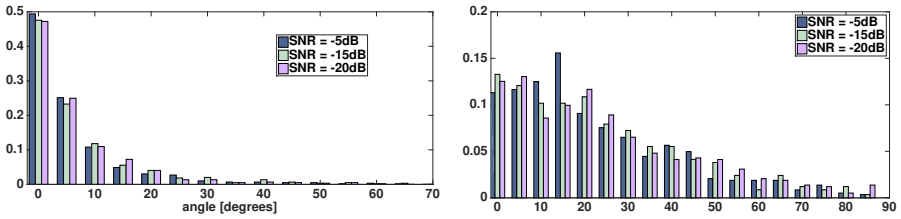
### 4.4 Experiment 3 and 4: Speaker Localisation

We perform speaker horizontal localisation at different SNR levels ($SNR \in \{-5dB, -15dB, -20dB\}$), comparing the behaviour of the *LOC-STEREO* and *LOC-CROSS* networks, when operating on frames of either $1s$ or $500ms$. We also compare the results obtained by both deep learning approaches with a more traditional geometric method based on Interaural Time Difference [19], indicated as *BASELINE*. In all cases, localisation of the target speaker is carried out against different noise combinations, as described in Section 4.1. Table 2 reports the results of those experiments. We observe that by employing either the *LOC-STEREO* or the *LOC-CROSS* networks, localisation accuracy degrades when dealing with shorter frames. This might be due to the fact that, since we do not apply any voice activity detection algorithm neither in training or testing, we might get more $500ms$-frames with not enough speech information for the network to train a reliable model. Indeed, the performance of the *BASELINE* approach, which doesn't rely on a specific learned model, greatly drop at lower SNRs, but shows consistent results independently on the frame size. We also see that using a stereo gammatonegram (*i.e.* gammatonegrams from both channels put side by side) remarkably help the localisation process, with respect to using the representation based on cross-correlation. We believe this is due to the presence of heavy noise, which might mask cues useful for speaker localisation when computing the cross-correlation. The difference between the behaviour of the two networks is reported in Figures 4 and 5. The greatest performance is obtained when using the *LOC-STEREO* on frames of $1s$, which provides a median error lower than 6° even when $SNR = -20dB$. Considering that the frames are randomly selected, and that, in reality, those will come as a consecutive audio

| | Localisation Accuracy | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **LOC-STEREO** | | | | **LOC-CROSS** | | | | **BASELINE** | | | |
| FL | $-5dB$ | $-15dB$ | $-20dB$ | All | $-5dB$ | $-15dB$ | $-20dB$ | All | $-5dB$ | $-15dB$ | $-20dB$ | All |
| 500 ms | 16.47 | 21.01 | 21.90 | 18.45 | 21.02 | 22.7 | 21.75 | 22.10 | 6.6 | 17.8 | 26.5 | 15.9 |
| 1 s | **5.09** | **5.32** | **5.43** | **5.24** | 19.93 | 22.38 | 23.4 | 21.57 | 5.5 | 16.7 | 27.1 | 15.3 |

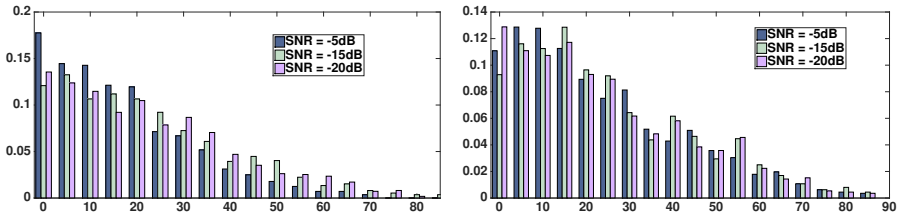
Table 2: The table reports the median of the the absolute error (expressed in degrees) in the DoA prediction obtained by using the *LOC-STEREO*, the *LOC-CROSS* networks, and the *BASELINE* approach. $FL$ refers to the frame length.

stream, the error could be further decreased through outlier removal (*i.e.* by applying a median filter).




(a) LOC-STEREO (b) LOC-CROSS

Fig. 4: From left to right: histogram of the absolute error in the localisation, when using the *LOC-STEREO* and the *LOC-CROSS* networks, operating on frames of $1s$.




(a) LOC-STEREO (b) LOC-CROSS

Fig. 5: From left to right: histogram of the absolute error in the localisation, when using the *LOC-STEREO* and the *LOC-CROSS* networks, operating on frames of $500ms$.

## 5 Conclusions

In this work, we explored speaker verification and horizontal localisation in challenging indoor and outdoor acoustic scenarios characterised by the presence of copious and unpredictable noise. We addressed both tasks employing a fully data-driven approach, based on CNN architectures. Our experimental evaluation, implemented on a robotic platform, demonstrated that the framework presented is able to perform both tasks (*i.e.* verification and localisation) robustly

and with a high level of accuracy. Future work will investigate the possibility of developing multi-modal systems (*e.g.* audio-visual) to enable more robust and accurate in-noise human-robot interaction. Furthermore, domain adaptation techniques could be used to reduce the amount of data necessary to train the models to extend the use of this framework to situations where only few speech data from the target speaker is available (*e.g.* a guide robot interacting with people in museums, airports or malls.)

## References

1. Abadi, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), `http://tensorflow.org/`
2. Bhattacharya, G., Alam, M.J., Kenny, P.: Deep speaker embeddings for short-duration speaker verification. In: Interspeech. pp. 1517–1521 (2017)
3. Chakrabarty, D., Elhilali, M.: Abnormal sound event detection using temporal trajectories mixtures. In: 2016 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 216–220 (2016)
4. Chan, W., Jaitly, N., Le, Q., Vinyals, O.: Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. pp. 4960–4964 (2016)
5. Deng, S., Han, J., Zhang, C., Zheng, T., Zheng, G.: Robust minimum statistics project coefficients feature for acoustic environment recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. pp. 8232–8236 (2014)
6. Feng, L.: Speaker Recognition. Ph.D. thesis, Technical University of Denmark, IMM-THESIS, DK-280, Kgs. Lyngby, Denmark (2004)
7. He, W., Motlicek, P., Odobez, J.M.: Deep neural networks for multiple speaker detection and localization. arXiv preprint arXiv:1711.11565 (2017)
8. Heigold, G., Moreno, I., Bengio, S., Shazeer, N.: End-to-end text-dependent speaker verification. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5115–5119 (2016)
9. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., Kingsbury, B.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine **29**(6), 82–97 (2012)
10. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
11. Holdsworth, J., Nimmo-Smith, I., Patterson, R., Rice, P.: Implementing a gammatone filter bank. Annex C of the SVOS Final Report: Part A: The Auditory Filterbank **1**, 1–5 (1988)
12. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Lyon, R.F., Katsiamis, A.G., Drakakis, E.M.: History and future of auditory filter models. In: Proceedings of 2010 IEEE International Symposium on Circuits and Systems. pp. 3809–3812 (2010)

14. Ma, N., May, T., Brown, G.J.: 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 (TASLP) **25**(12), 2444–2453 (2017)
15. Maganti, H.K., Matassoni, M.: Auditory processing inspired robust feature enhancement for speech recognition. In: International Joint Conference on Biomedical Engineering Systems and Technologies. pp. 205–218 (2011)
16. Marchegiani, L., Fafoutis, X.: On cross-language consonant identification in second language noise. The Journal of the Acoustical Society of America **138**(4), 2206–2209 (2015)
17. Marchegiani, L., Newman, P.: Learning to listen to your ego(-motion) : Metric motion estimation from auditory signals. In: Towards Autonomous Robotics Systems (TAROS). pp. 247–259 (2018)
18. Marchegiani, L., Posner, I.: Leveraging the urban soundscape: Auditory perception for smart vehicles. In: Robotics and Automation (ICRA), 2017 IEEE International Conference on. pp. 6547–6554 (2017)
19. Marchegiani, M.L., Pirri, F., Pizzoli, M.: Multimodal speaker recognition in a conversation scenario. In: International Conference on Computer Vision Systems. pp. 11–20 (2009)
20. Noda, K., Hashimoto, N., Nakadai, K., Ogata, T.: Sound source separation for robot audition using deep learning. In: Humanoid Robots (Humanoids), 2015 IEEE-RAS 15th Int Conf on. pp. 389–394 (2015)
21. Reynolds, D.A.: An overview of automatic speaker recognition technology. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 4, pp. IV–4072 (2002)
22. Rudzyn, B., Kadous, W., Sammut, C.: Real time robot audition system incorporating both 3d sound source localisation and voice characterisation. In: Robotics and Automation, 2007 IEEE International Conference on. pp. 4733–4738 (2007)
23. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: 22st ACM International Conference on Multimedia (ACM-MM'14). Orlando, FL, USA (Nov 2014)
24. Schluter, R., Bezrukov, I., Wagner, H., Ney, H.: Gammatone features and feature combination for large vocabulary speech recognition. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. vol. 4, pp. IV–649 (2007)
25. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: Interspeech. pp. 999–1003 (2017)
26. Stefanov, K., Sugimoto, A., Beskow, J.: Look who's talking: visual identification of the active speaker in multi-party human-robot interaction. In: Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction. pp. 22–27. ACM (2016)
27. Takahashi, N., Gygli, M., Van Gool, L.: Aenet: Learning deep audio features for video analysis. arXiv preprint arXiv:1701.00599 (2017)
28. Tapus, A., Bandera, A., Vazquez-Martin, R., Calderita, L.V.: Perceiving the person and their interactions with the others for social robotics–a review. Pattern Recognition Letters **118**, 3–13 (2019)
29. Toshio, I.: An optimal auditory filter. In: Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on. pp. 198–201 (1995)