Why talk to people when you can talk to robots? Far-field speaker identification in the wild

Humblot-Renaux, Galadrielle; Li, Chen; Chrysostomou, Dimitrios

Published in:
2021 30th IEEE International Conference on Robot and Human Interactive Communication, RO-MAN 2021

DOI (link to publication from Publisher):
10.1109/RO-MAN50785.2021.9515482

Publication date:
2021

Document Version
Early version, also known as pre-print

Why talk to people when you can talk to robots? Far-field speaker identification in the wild

Galadrielle Humblot-Renaux, Chen Li, Dimitrios Chrysostomou

***Abstract*— Equipping robots with the ability to identify who is talking to them is an important step towards natural and effective verbal interaction. However, speaker identification for voice control remains largely unexplored compared to recent progress in natural language instruction and speech recognition. This motivates us to tackle text-independent speaker identification for human-robot interaction applications in industrial environments. By representing audio segments as time-frequency spectrograms, this can be formulated as an image classification task, allowing us to apply state-of-the-art convolutional neural network (CNN) architectures. To achieve robust prediction in unconstrained, challenging acoustic conditions, we take a data-driven approach and collect a custom dataset with a far-field microphone array, featuring over 3 hours of "in the wild" audio recordings from six speakers, which are then encoded into spectral images for CNN-based classification. We propose a shallow 3-layer CNN, which we compare with the widely used ResNet-18 architecture: in addition to benchmarking these models in terms of accuracy, we visualize the features used by these two models to discriminate between classes, and investigate their reliability in unseen acoustic scenes. Although ResNet-18 reaches the highest raw accuracy, we are able to achieve remarkable online speaker recognition performance with a much more lightweight model which learns lower-level vocal features and produces more reliable confidence scores. The proposed method is successfully integrated into a robotic dialogue system and showcased in a mock user localization and authentication scenario in a realistic industrial environment: `https://youtu.be/IVtZ8LKJZ7A`.**

I. Introduction

Auditory perception for intelligent systems has shown promising applications ranging from alarming event detection [1], to object recognition [2] and ego-motion estimation [3]. It especially plays a central role in industrial and social robotics by allowing users to communicate with robots through speech [4], [5]. Audio processing and modelling techniques can be used not only to identify *what* the user is saying (speech recognition), but also *where* they are (sound source localisation), and *who* they are (speaker recognition). When tackled as a *verification* task, speaker recognition only involves a binary decision for authentication; in contrast, speaker *identification* is a multi-class problem which bears more relevance to human-robot interaction (HRI) as it allows us to personalize responses based on who is talking [6]. While speech recognition has gained widespread use due to publicly available APIs and pre-trained models which are able to generalize to new voices [7], speaker identification is more challenging to apply beyond offline benchmarks as it requires collecting diverse data for each user that we wish to identify, and thus remains largely unexplored in a robotic context. We highlight the untapped potential of speaker identification to enrich HRI, as it can allow a robot to partake in multi-user scenarios and adapt its responses based on the identity and competences of each speaker.

As part of our ongoing research on integrating virtual assistants in industrial robots [8], we consider the use case of an autonomous industrial mobile manipulator (AIMM) operating in an industrial setting, equipped with a speaker and a far-field microphone array for verbal interaction with multiple users. Such a set-up brings significant challenges for speech processing, as the robot operates in an unpredictable acoustic environment containing diverse sound sources, ambient noise and reverberation. Besides, the user's variable distance can significantly affect audio quality and recognition performance [8], [9], while noise generated by the robot's motion further degrades speech wave-forms [7].

As illustrated in Figure 1, our goal is to incorporate speaker identification into a real-world dialogue system, where a robot assisting shop floor workers in their daily work is able to identify known speakers from their voice commands, address them by name and turn to their direction. This requires real-time, robust classification of audio samples, despite ambient noise, speaker distance and robot movement. To this end, we propose a simple yet effective CNN-based approach to recognize six different speakers and also identify samples containing no speech. Unlike works which artificially degrade clean samples to evaluate performance at different Signal-to-Noise Ratio (SNR) levels [10]–[14], we develop and evaluate our method on a custom dataset captured in diverse, realistic and noisy indoor environments. Although this approach makes it more

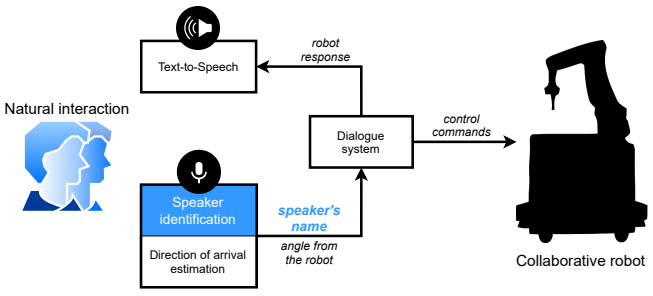


Fig. 1. Overview of the proposed interactive system. Our main contribution is the speaker identification method (highlighted in blue), which we first develop and evaluate separately, and then integrate into a real-world collaborative robotic application on the Little Helper 8 platform.

Authors are members of the Robotics & Automation Group, Department of Materials and Production, Aalborg University, Denmark, ghumbl19@student.aau.dk; {cl,dimi}@mp.aau.dk

difficult to control and characterize the audio quality, we posit that it produces models with greater generalization capability and better indicates true performance during real-life operation. In addition, while most existing works only perform offline evaluation in terms of accuracy, we take a comprehensive evaluation approach, where we investigate the model's reliability, visualize learned features with Class Activation Maps (CAMs), and demonstrate its real-world use through an interactive system developed on our latest AIMM iteration, Little Helper 8 [15], [16].

In the remainder of this paper, we first identify recent related work in Section II and outline our proposed method for spectrogram generation and CNN-based classification in Section III. We then describe the data collection procedure for extensive evaluation of the model and present a fully-integrated speaker-aware robotic system in Section IV. Lastly, Section V summarizes the main results and findings and addresses some limitations of the current implementation, leading to promising directions for future work.

II. Related Work

A. Audio representation for classification

Time-frequency representation of audio signals facilitates visual interpretation and classification. Rather than representing the frequency spectrum on a linear scale, a widespread speech processing technique consists of applying filter banks along a warped frequency scale in a perceptually meaningful way, with higher resolution in the low-frequency range to mimic human audition. Frequency-warped spectrogram representations have repeatedly been shown to out-perform frequency cepstral coefficients (e.g. MFCCs) and linear-scale spectrograms in a wide range of CNN-based audio classification tasks [1], [17]–[20].

While mel-filtered spectrograms are frequently the feature representation of choice for speaker recognition [14], [17], [21], existing work tackling audio classification for robotics [1], [20] and speaker verification in low SNR conditions [13] suggests that robust classification can also be achieved by applying gammatone filters to obtain "gammatonegrams". Thus, we compare both input representations in our experiments.

B. CNN-based speaker identification

CNNs have shown to be a promising alternative to traditional methods such as the Gaussian mixture model (GMM) for learning speaker-specific features [6]. For instance, [21] presents a light-weight CNN architecture for mel-spectrogram classification, achieving remarkable performance on the TIMIT [22] dataset, consisting of extremely clean audio. The Interspeech speaker recognition challenges and release of "in the wild" datasets collected from open-source media such as Speakers in the Wild [23] and VoxCeleb [24] has sparked significant progress in deep, robust text-independent speaker identification. Many state-of-the-art works [17], [18], [25] on these two benchmarks employ a ResNet-based [26] architecture, which has been widely adopted for general image classification. However, the neural architecture search approach followed in [27] suggests that the model complexity offered by this widely used back-bone may not be necessary for speaker identification performance and that developing task-specific neural architectures yields significant gains. Recent work investigating neural network calibration [28] also suggests that the increase in accuracy brought by modern deep neural networks such as ResNet may come at the expense of reliable confidence scores, which is a concern for robotic applications where predictions are used as a basis for decision-making and control. This motivates us to investigate the reliability of ResNet for speaker identification compared to a shallower architecture.

C. Robot audition for user identification

We highlight the lack of existing closed-loop, deep learning-based, and noise-robust systems: most existing works tackling speaker recognition for HRI do not consider adverse factors such as ambient noise and speaker distance, only use a robotic platform for data collection but not for closed-loop inference, or employ traditional methods. For instance, [29] presents a real-time speaker verification system for HRI based on hand-crafted features and pattern matching and validates the method on clean samples from a public voice database. More recently, [30] trains GMMs to classify five speakers and demonstrates how speaker identification could be incorporated in a real system for voice control of a humanoid robot with a microphone array - however, close-range interaction with pre-defined commands and ideal acoustic conditions are assumed. While [12] specifically addresses robustness to noise, background chatter and distance for in-vehicle speaker identification; it also employs traditional classification methods and relies on simulated environmental conditions for evaluation.

Similarly to our approach, [13] records speech from three speakers with a moving robot, and investigates the effect of additive noise for CNN-based spectral image classification, showing that high accuracy can be achieved with a shallow network despite low SNR and short sample length. However, [13] tackles speaker *verification* rather than identification and is not evaluated beyond offline performance on artificially degraded samples which were recorded in a clean acoustic environment. Furthermore, only a single architecture and spectral representation are considered.

In contrast, our work demonstrates online inference in a robotic application under challenging, unconstrained conditions while applying state-of-the-art deep learning methods rather than hand-crafted features. We compare two spectral image representations and two CNN architectures (a 3-layer network against the smallest 18-layer variant of ResNet) and investigate their performance for real-world speaker identification in terms of accuracy, reliability and efficiency.

III. Methods

The main idea of the proposed approach is illustrated in Figure 2. Given a raw audio stream, we first extract three fixed-length segments, as described in Section III-A. In Section III-B, we explain how each audio segment is transformed

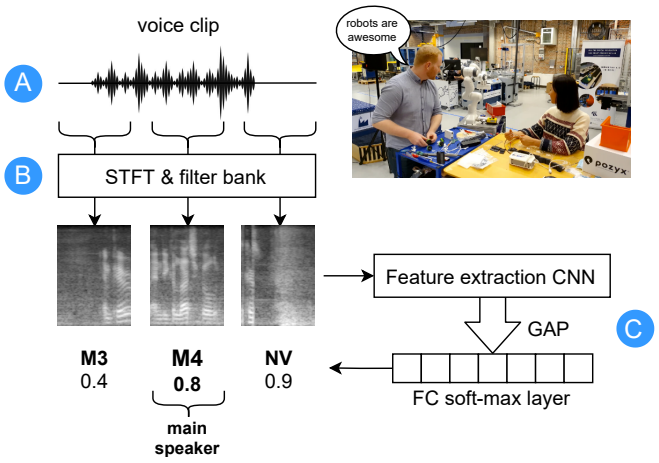


Fig. 2. Overview of our 3-stage speaker identification method, from audio recording to final prediction, with the corresponding sub-sections indicated in blue. M1 and M4 correspond to two speakers in our dataset, and NV refers to the no voice class. The picture in the top right shows the experimental set-up used for data collection and testing.

into a time-frequency spectral image by generating its filtered and frequency-warped Short-time Fourier Transform (STFT). For feature extraction and classification, the resulting images are fed to a CNN, which we train to recognize a small set of speakers. As presented in Section III-C, its architecture is based on Global Average Pooling (GAP) and features a single fully-connected (FC) output layer. This yields three predictions for each audio clip; the main speaker is extracted based on class probabilities across three audio segments.

A. Generating samples for classification

As input to our system, we consider a stream of audio recorded with a ReSpeaker far-field 4-microphone array and sampled at $f_s = 16\,\text{kHz}$. The device provides built-in voice activity detection (VAD), which is applied as a pre-processing step to extract voice clips. The VAD method is energy-based, thus triggered not only by voice but also other acoustic events, e.g. mouse scroll and coughing, which is why we introduce an additional *no voice* (NV) class in our dataset (cf. Section IV-B). Although VAD is not necessary in our set-up, it avoids having to continuously classify samples in the absence of acoustic activity. Since a sliding window approach does not scale well with long utterances in terms of computational load, we only generate three 1-second samples per voice clip, regardless of its length: a sample is extracted at the beginning, middle and end of the voice clip - as illustrated in Figure 2. The short audio sample length, while detrimental to speaker recognition performance [25], allows us to classify audio on-the-fly due to the small input size.

B. Encoding audio into spectral images

Each 1-second audio sample is converted to a spectral image, which encodes the signal energy for a given time and frequency as pixel intensity. Similarly to [20] and [18], we specifically apply and compare two different filter banks in our experiments, introduced in Section II: *mel* filters (triangular bandpass filters applied on a mel-scale) and *gammatone* filters (asymmetrical linear bandpass filters which aim to approximate the human auditory frequency response, applied on an Equivalent Rectangular Bandwidth scale).

In both cases, as [21], a power spectrum is first generated by applying an STFT on the audio sample over a 1024-point Hann window (giving a frequency resolution of 16 Hz) with a hop length of 160 (10 ms). A 128-channel filter bank is then applied on the spectrum, with a min. frequency of 20 Hz and max. frequency of $f_s/2$. The power spectrum is then converted to a logarithmic scale. An example of the resulting spectrum for the two filter banks are shown in Figure 3.

Lastly, an image is obtained by shifting & scaling the spectrogram matrix to fit in an 8-bit range of 0-255. This yields images of sizes 128×94 for 1-second audio samples. We also investigated the effect of applying cepstral mean and variance normalization as recommended by [17], [24], however, we found that it degrades classification performance in our set-up, therefore, we omit this step.

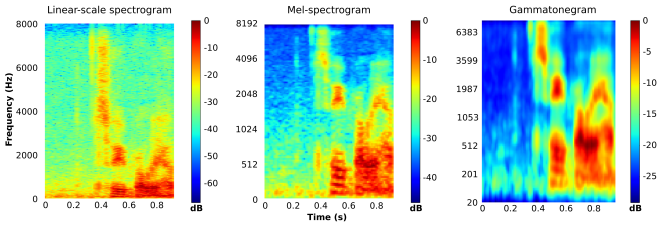


Fig. 3. Color visualization of log-power spectrograms generated for an input sample, comparing different frequency scaling methods. The linear-scale spectrogram is the output of the STFT. This sample contains speech from 0.38 to 1 second

C. Convolutional neural network architecture

In order to learn discriminative features from spectral images, we propose a shallow CNN, which we call *SpeakerNet* and illustrate its architecture in Figure 4. The feature extraction stage takes a single-channel image as input and follows a generic structure, consisting of three convolutional layers with ReLU activation. Based on [13], we use a convolutional kernel size of 5×5 with zero-padding to preserve spatial resolution, and 2×2 max-pooling of stride 2 for down-sampling. Batch normalization and drop-out layers (rate of 0.1) are added for regularization.

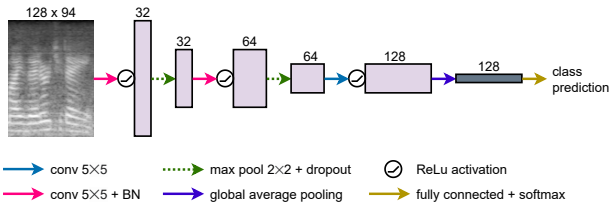


Fig. 4. The architecture of the proposed SpeakerNet network

Since we are interested in visualizing learned representations, we employ the method described in [31] to generate CAMs, which offer qualitative insight into which regions in the spectrograms are the most relevant to recognize each class in our dataset. This requires applying GAP after the

feature extraction stage, with a single final FC layer for classification. As opposed to the spectrogram classification architectures in [13], [19], [21], for instance, discarding intermediate fully-connected layers in favor of GAP has the added benefit of parameter reduction and regularization.

We compare SpeakerNet with a ResNet-18 [26] architecture, which is used without any alteration except for the size of the fully-connected output layer to match the number of classes in our dataset. The model is initialized with pre-trained weights from ImageNet [32]. The architecture already features GAP and thus can be directly used for generating CAMs. Since the network was designed for 3-channel image input, we repeat the same gray-scale spectrogram across three dimensions.

IV. Experiments

A. Scenario description

We showcase the system in a noisy industrial environment, following two different scenarios. To mimic a user authentication application, and in a similar spirit to [30], the identity of the detected speaker is checked before executing any control commands - if an impostor is recognized, the robot does not turn and notifies them that they are unauthorized to give commands. In the first scenario (shown in Figure 2), we simulate a collaborative manufacturing arrangement: two speakers are working on an assembly task, with the robot in their vicinity ($\sim$ 1 m), however only one of them is authorized to issue robot instructions. The impostor first tries to request a tool from the robot and then asks her co-worker to repeat the voice command. In the second scenario, we test the model's ability to recognize a wider variety of voices, and at a greater distance: the four speakers stand at $\sim 2\,\mathrm{m}$ from the robot, and in turn, ask it to turn towards them.

B. Data collection

While large-scale speaker recognition benchmarks feature hundreds of voices recorded with different devices [23], [24], we aim to capture high intra-speaker variability from six users with a fixed recording set-up. Rather than setting up a controlled experiment where each speaker is recorded separately and given specific instructions (e.g. as in [12], [22], [30]), we capture raw multi-speaker audio "in the wild". The dataset is recorded across ten sessions over the course of two weeks, with the microphone array placed in different rooms with various acoustic conditions, i.e. a silent office, a busy research lab, a noisy industrial shop floor, a chatty break room and calm corridors capturing a combination of natural conversations, meetings, and human-to-robot commands (e.g. "Hey robot, can you give me the screwdriver?"). No constraints are enforced on the recording content or quality: collected samples feature real-world noise, including unintelligible background chatter, laughter, machinery noise, and keyboard tapping with speakers standing at various distances ($\approx$ 20cm to 5m) and directions to the microphone, talking freely in a variety of tones (e.g. mumbling, excited, amused, or hesitant). Incorporating such variation in vocal style, and spacing out recording sessions over time, has shown to bring added robustness when training speaker recognition models [33].

In the data collection stage, to match real operation conditions, VAD clips are segmented into 1-second samples, saved on-the-fly and manually sorted at a later stage by listening to each sample to determine the identity of the speaker. Segments containing no discernible speech are assigned to a *no voice* (NV) class (this includes background chatter, silence, and any noise). Segments with more than one speaker in the foreground are discarded. As a result, we have a 7-class dataset with an NV class and six voices: A1 (androgynous), F1 (female) and M{1,2,3,4} (male speakers), adding up to a total of 11202 1-second audio samples (over 3 hours).

Table I shows examples of spectral images generated from challenging samples for each class: the NV class is the easiest to discern due to the absence of vocal features, and the female speaker F1 has a high-pitched voice characterized by widely spaced vocal lines in the spectrograms. However, the five other voices are difficult to distinguish visually from spectral images, especially given the diversity in utterance tone, content and recording conditions. Many samples contain minimal vocal content over time, coupled with interference from background noise and chatter as well as sharp, loud sounds.

C. Training procedure

For both architectures, the model is trained over a maximum of 600 epochs by optimizing cross-entropy loss via Stochastic Gradient Descent with Warm Restarts (SGDR) [34] and momentum of 0.9 to accelerate convergence. For regularization, a weight decay factor of 0.001 is applied. Input samples are normalized based on the mean and standard deviation of the training set and fed to the model in batches of size 64.

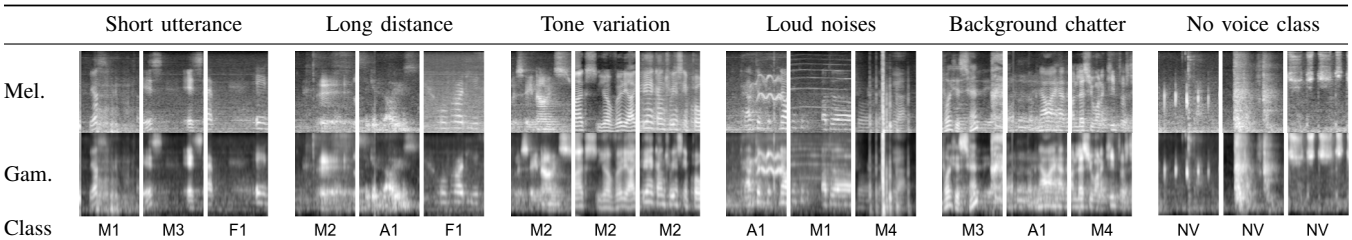
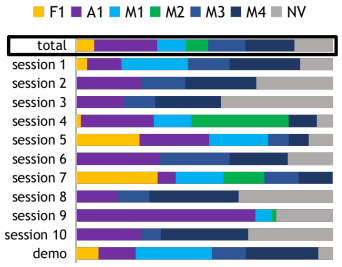


TABLE I. Different types of challenging samples, showing the diversity in our custom dataset. Each sample corresponds to an audio segment of exactly 1 second, which we convert to mel-spectrogram (top) and gammatonegram (bottom) images for training the CNN models.



Split	Scene	Mel. ResNet-18	Mel. SpeakerNet	Gam. ResNet-18	Gam. SpeakerNet
10%	5 random splits, mixed scenes, all speakers	**95.45 ± 0.45**	**91.22 ± 2.74**	95.29 ± 0.34	89.72 ± 3.02
1	shop floor, commands & conversation, variable distances, ambient noise	90.07	79.66	**90.35**	**80.13**
2	lab, commands & conversation, variable distances with loud sounds	92.36	**87.18**	**92.8**	84.73
3	lab, commands & conversation, close range with loud sounds	92.00	85.89	**93.05**	**87.58**
4	meeting room & shop floor, conversation, variable distances, background chatter	**54.16**	48.00	53.08	**48.56**
5	break room, conversation, close range with loud sounds, clean background	74.96	**68.50**	**75.76**	63.64
6	break room & shop floor, conversation, variable distances with loud sounds, chatter & noise	**86.89**	82.40	86.32	**83.00**
7	shop floor, reading text, far, ambient noise	68.34	61.08	**74.77**	**62.94**
8	shop floor, conversation, close range with loud sounds, chatter & noise	81.63	83.67	**87.76**	**87.76**
9	meeting room, close range with loud sounds, clean background	97.54	90.98	**98.36**	**91.8**
10	lab, commands, variable distances, robot movement and other noises	**95.87**	**85.12**	95.04	**85.12**
demo	shop floor, cf. Section IV-H	**89.74**	**82.65**	88.25	78.36

Fig. 5. Proportion of samples per speaker in the full dataset and for individual sessions (left). The table gives the accuracy (percentage) per train/validation split, with the best input representation for each model is highlighted in bold. Note that due to the unconstrained recording procedure, the description is not necessarily representative of every sample in a session, but rather aims to capture the general features of the scene.

D. Classification accuracy

To get a baseline for how well the two architectures can discriminate between the 7 classes in our dataset, we first perform random cross-validation, such that both the training and validation set contain samples across all recording sessions. To prevent any overlap between the training and validation set, segments from the same audio clip are assigned to the same set. The dataset is split into five folds with a train/validation split of 90/10% in each fold. The mean validation accuracy and standard deviation across the five folds are given in the first row of the table in Figure 5. In this experiment, ResNet-18 achieves over 95% accuracy while SpeakerNet approaches 90% for both input representations. We also find that ResNet-18 is much more prone to over-fitting than our shallow model, especially for the mel-spectrogram representation.

In order to assess the models' ability to generalize to new recording sessions, we then split the dataset by date, such that a recording session is left out as the validation set, and other sessions are used for training. For each split, the validation accuracy is reported in the rest of the table. ResNet-18 consistently reaches higher accuracy than SpeakerNet except in session 8, which contains the least amount of samples from all validation sets. This suggests that given more training data, SpeakerNet may become more accurate. Both models struggle the most in sessions 4 and 7, which are particularly challenging: session 4 contains a large portion of samples from speaker M2 (green in Figure 5) who was absent in most other recording sessions, thus leaving minimal training data for the model, and in session 7 the users read a text in a monotone voice, which contrasts from the rest of the dataset which primarily features lively, natural speech.

We observe that the classes which the highest occurrence in the dataset (A1, M4 and NV) also yield the highest accuracy for both models. Thus, aligning with the concerns raised in [33], we highlight the importance of recording substantial and diverse voice data for every user in order to achieve robust speaker identification in new situations. Lastly, we find that the preferred input representation varies across sessions and per model: upon inspection of misclassified samples, it seems that gammatonegrams provide higher robustness when the speech is masked by loud noises (aligning with the findings in [13]), while mel-spectrograms are able to yield higher accuracy when uttering clean commands. Looking at the accuracy for mel./gam. spectrograms across all cross-session validation samples, ResNet-18 reaches 78.85/**79.37**% as opposed to **72.61**/72.50% for SpeakerNet.

E. Feature visualization with CAMs

We generate CAMs for all cross-session validation samples as described in Section III-C - a few examples for correct and incorrect predictions by the two models are shown in Figure II. As expected, due to its significantly narrower receptive field, SpeakerNet is activated by much finer features than ResNet-18. Interestingly, for SpeakerNet, the most informative regions for speaker identification lie towards the lower end of the spectrum, around the fundamental frequency and first harmonics - with different activation patterns for gammatonegrams and mel-spectrograms. SpeakerNet's confusion when making incorrect predictions frequently transpires through the CAMs, with high activation across the whole spectrum.

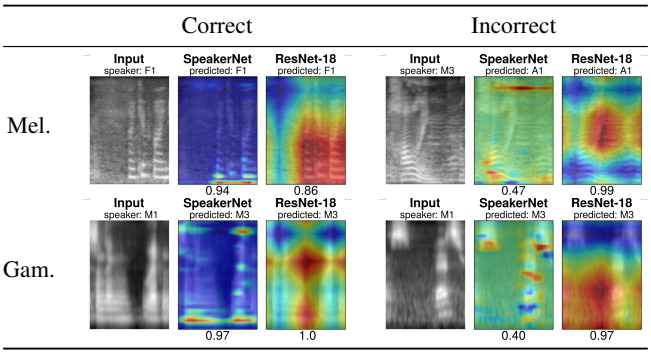


TABLE II. CAMs generated from predictions on unseen samples, with the most informative regions in red and least informative regions in blue. The confidence score is also shown under each prediction.

F. Neural network reliability

We evaluate the models' reliability in terms of Expected Calibration Error (ECE) and Maximum Calibration Error (MCE), as defined in [28]. This gives a measure of whether the probability of the predicted class is a reliable indicator of expected accuracy. Ideally, the more confident a model is in its prediction, the higher the likelihood that the prediction is correct, and vice-versa - this allows us to determine whether a prediction should be trusted or ignored.

As shown in Figure 6, ResNet-18 tends to deliver over-confident predictions, with the majority of both correct and incorrect predictions having a confidence > 0.9, while SpeakerNet produces significantly better calibrated scores. While the reliability plot is only shown for mel-spectrogams, we observe similar results for gammatonegrams, with an ECE/MCE of 14.14/31.40 for ResNet-18 and 6.91/9.10 for SpeakerNet. In our set-up, reliable confidence scores provide robustness to misclassification errors since for each voice clip, we pick the highest-confidence prediction out of 3 samples to identify the speaker.

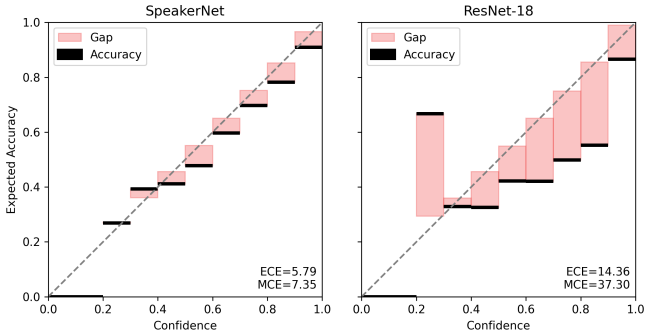


Fig. 6. Reliability diagrams for both models, generated from mel-spectrogram predictions across all the session validation sets in Table 5. The dotted line indicates the expected accuracy for a perfectly calibrated model (ECE and MCE of 0). Error metrics are shown on the bottom right.

G. Computational footprint

We benchmark the two models on our hardware set-up, where inference is performed on a high-end laptop running Ubuntu 18.04, equipped with an Intel Core i7-6820HQ CPU @ 2.70GHz and 64GB RAM. As shown in Table III, due to the small input size (image resolution of 128×94), both models are suitable for on-board speaker identification. However, SpeakerNet is able to classify samples at over double the speed of ResNet-18.

	ResNet-18	SpeakerNet
trainable parameters	11,180,103	**258,119**
speed (ms)	22	**9**

TABLE III. Comparison of the models' footprint. The speed is measured as the average forward pass time for a batch of 3 samples.

H. Integration into a robotic system

We deploy our interactive system on a Little Helper 8 presented in [8], consisting of a Franka Emika Panda collaborative arm mounted on a MiR200 platform for autonomous navigation. We fix the microphone on the robot's table, and run the full system on the hardware described in Section IV-G (with no GPU acceleration), with the laptop's speaker used as the robot's voice. For capturing live audio, we use the same microphone as in the data collection stage. The device includes built-in direction of arrival (DoA) estimation, which provides the person's approximate position relative to the robot and rotates it towards the speaker. A Text-to-Speech (TTS) engine is used to generate robotic responses based on the speaker's identity, e.g. "Okay, Chen, I will move towards your direction". To avoid the robot's voice being classified or the ongoing task being interrupted, we ignore audio activity when the TTS output is playing, or the robot moves. For online speaker identification, based on the results reported in Section IV, we select SpeakerNet as the most light-weight and reliable model. Mel-spectrogram images are generated and classified on-the-fly in sets of 3 from VAD audio clips following the same procedure as in the data collection stage and illustrated in Figure 2. The final prediction is taken as the class with the highest confidence, excluding the NV class.

A video of both tests is available at `https://youtu.be/IVtZ8LKJZ7A`, featuring four of the six speakers in our dataset. Despite the high level of ambient noise from ventilation and machinery, distance to the robot, and clutter around the microphone (in the first scenario), each speaker is correctly identified in both tests; the robot responds immediately and rotates towards each authorized speaker. We also collected and labelled the samples of this demo session for offline evaluation - results are shown in the last row of Figure 5. We found that audio segments containing particularly brief utterances are frequently misclassified; however, they are successfully excluded from the final speaker prediction due to their low confidence score.

V. Discussion and Conclusion

We have shown how state-of-the-art CNN architectures can be applied for online speaker identification in noisy conditions and used as part of an interactive robotic dialogue system. We compare two CNN models and two filter banks in a particularly challenging set-up, with short input samples, high SNR variability and diverse vocal content. Our results demonstrate that SpeakerNet has a significantly lower computational footprint and learns lower-level features than ResNet-18, yet generalizes well to new recording sessions in challenging conditions while producing much more reliable confidence scores, which is crucial in our set-up where we use these scores to extract the main speaker from predictions on a set of short fixed-length audio segments.

Our "in the wild" data collection set-up yields realistic audio samples which resemble those encountered during normal operation and allows the tested CNN models to learn sufficiently general features for recognizing speakers across new recording sessions in noisy environments. However, the resulting raw, jumbled data makes it difficult to characterize the effect of specific recording conditions on model performance and a more systematic evaluation of noise level, speaker distance and vocal characteristics would provide further insight. Furthermore, manually sorting hundreds of 1-second clips for training the model is tedious and time-consuming. Exploring weakly supervised approaches would be beneficial to reduce the effort of manual labelling.

Additionally, a significant limitation of our implementation is that it assumes a close-set of possible voices, with each input sample only featuring the voice of a single person in the foreground. In practice, recording voice activity in

a natural environment often results in segments containing speech from unknown speakers, or multiple people talking in close succession or at the same time. Speech separation and metric learning for open-set speaker identification are outside of the the scope of this work, but are an important direction for future work.

Lastly, in a robotic context where different sensors are often already available, incorporating visual representations instead of solely relying on audio for user identification would be highly beneficial. Coupling these two modalities may also allow us to distinguish between speech which is directed to the robot, as opposed to other people in the scene.

ACKNOWLEDGEMENTS

We would like to thank the research assistants Martin Bieber, Jinha Park and Hahyeon Kim for their significant help with robotic integration and lending their voices to this experiment. We would also like to thank Letizia Marchegiani for her valuable technical input. Lastly, we would like to acknowledge support by the H2020-WIDESPREAD project no. 857061 "Networking for Research and Development of Human Interactive and Sensitive Robotics Taking Advantage of Additive Manufacturing – R2P2".

REFERENCES

[1] L. Marchegiani and I. Posner, "Leveraging the urban soundscape: Auditory perception for smart vehicles," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 6547–6554.

[2] D. Gandhi, A. Gupta, and L. Pinto, "Swoosh! rattle! thump! – actions that sound," 2020.

[3] L. Marchegiani and P. Newman, "Learning to listen to your ego-(motion): Metric motion estimation from auditory signals," in *Towards Autonomous Robotics Systems (TAROS)*, 2018, p. 247–259.

[4] M. C. Bingol and O. Aydogmus, "Performing predefined tasks using the human–robot interaction on speech recognition for an industrial robot," *Engineering Applications of Artificial Intelligence*, vol. 95, p. 103903, 2020.

[5] A. Tapus, A. Bandera, R. Vazquez-Martin, and L. V. Calderita, "Perceiving the person and their interactions with the others for social robotics – a review," *Pattern Recognition Letters*, vol. 118, pp. 3 – 13, 2019.

[6] R. Mohd Hanifa, K. Isa, and S. Mohamad, "A review on speaker recognition: Technology and challenges," *Computers & Electrical Engineering*, vol. 90, p. 107005, 2021.

[7] J. Novoa, J. Wuth, J. P. Escudero, J. Fredes, R. Mahu, and N. B. Yoma, "Dnn-hmm based automatic speech recognition for hri scenarios," in *ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 2018, p. 150–159.

[8] C. Li, J. Park, H. Kim, and D. Chrysostomou, "How can i help you? an intelligent virtual assistant for industrial robots," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 2021, p. 220–224.

[9] M. A. Nematollahi and S. Al-Haddad, "Distant speaker recognition: An overview," *Int. J. Humanoid Robotics*, vol. 13, pp. 1 550 032:1–1 550 032:45, 2016.

[10] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 3997–4001.

[11] M. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, "A robust speaker identification system using the responses from a model of the auditory periphery," *PLoS ONE*, vol. 11, 2016.

[12] I. Bisio, C. Garibotto, A. Grattarola, F. Lavagetto, and A. Sciarrone, "Smart and robust speaker recognition for context-aware in-vehicle applications," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8808–8821, 2018.

[13] T. Tse, D. De Martini, and L. Marchegiani, "No need to scream: Robust sound-based speaker localisation in challenging scenarios," in *11th International Conference on Social Robotics (ICSR)*, ser. Lecture Notes in Computer Science, vol. 11876. Springer, 2019, pp. 176–185.

[14] A. M. Jalil, F. S. Hasan, and H. A. Alabbasi, "Speaker identification using convolutional neural network for clean and noisy speech samples," in *First International Conference of Computer and Applied Sciences (CAS)*, 2019, pp. 57–62.

[15] R. E. Andersen, E. B. Hansen, D. Cerny, S. Madsen, B. Pulendralingam, S. Bøgh, and D. Chrysostomou, "Integration of a skill-based collaborative mobile robot in a smart cyber-physical environment," *Procedia Manufacturing*, vol. 11, pp. 114–123, 2017.

[16] C. Schou, R. S. Andersen, D. Chrysostomou, S. Bøgh, and O. Madsen, "Skill-based instruction of collaborative robots in industrial settings," *Robotics and Computer-Integrated Manufacturing*, vol. 53, pp. 72–80, 2018.

[17] N. N. An, N. Q. Thanh, and Y. Liu, "Deep cnns with self-attention for speaker identification," *IEEE Access*, vol. 7, pp. 85 327–85 337, 2019.

[18] D. Cai, X. Qin, W. Cai, and M. Li, "The DKU System for the Speaker Recognition Task of the 2019 VOiCES from a Distance Challenge," in *INTERSPEECH*, 2019, pp. 2493–2497.

[19] A. Lieto, D. Moro, F. Devoti, C. Parera, V. Lipari, P. Bestagini, and S. Tubaro, ""hello? who am i talking to?" a shallow cnn approach for human vs. bot speech classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2577–2581.

[20] D. Williams, D. De Martini, M. Gadd, L. Marchegiani, and P. Newman, "Keep off the grass: Permissible driving routes from radar with weak audio supervision," in *IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–6.

[21] Y. Lukic, C. Vogt, O. Dürr, and T. Stadelmann, "Speaker identification and clustering using convolutional neural networks," in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.

[23] M. McLaren, L. Ferrer, D. Castán, and A. Lawson, "The speakers in the wild (sitw) speaker recognition database," in *INTERSPEECH*, 2016.

[24] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*, 2018.

[25] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 5791–5795.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[27] S. Ding, T. Chen, X. Gong, W. Zha, and Z. Wang, "AutoSpeech: Neural Architecture Search for Speaker Recognition," in *INTERSPEECH*, 2020, pp. 916–920.

[28] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 1321–1330.

[29] F. Alonso Martín, A. Ramey, and M. A. Salichs, "Speaker identification using three signal voice domains during human-robot interaction," in *ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, 2014, p. 114–115.

[30] I.-J. Ding and J.-Y. Shi, "Kinect microphone array-based speech and speaker recognition for the exhibition control of humanoid robots," *Computers & Electrical Engineering*, vol. 62, pp. 719 – 729, 2017.

[31] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[32] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[33] T. F. Zheng and L. Li, "Speaker-related robustness issues," in *Robustness-Related Issues in Speaker Recognition*. Springer, 2017, pp. 27–37.

[34] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *5th International Conference on Learning Representations, ICLR*, 2017.