RoboVox: A Single/Multi-channel Far-field Speaker Recognition Benchmark for a Mobile Robot

Mohammad Mohammadamini*, Driss Matrouf*, Michael Rouvier*, Jean-Francois Bonastre*, Romain Serizel†, Theophile Gonos‡

*Avignon University, †University of Lorraine, ‡AI Mergence

LIA (Laboratoire Informatique d'Avignon), CNRS, Inria, Loria, F-54000, Nancy, France,

Parv. Alan Turing, 75013 Paris

{mohammad.mohammadamini, driss.matrouf, michael.rouvier,

jean-francois.bonastre}@univ-avignon.fr, romain.serizle@loria.fr, theophile.gonos@ai-mergence.com

Abstract

In this paper, we introduce a new far-field speaker recognition benchmark called RoboVox. RoboVox is a French corpus recorded by a mobile robot. The files are recorded from different distances under severe acoustical conditions with the presence of several types of noise and reverberation. In addition to noise and reverberation, the robot's internal noise acts as an extra additive noise. RoboVox can be used for both single-channel and multi-channel speaker recognition. In the evaluation protocols, we are considering both cases. The obtained results demonstrate a significant decline in performance in far-filed speaker recognition and urge the community to further research in this domain.

Keywords: Speaker recognition, Far-field, Noise, Reverberation, RoboVox

1. Introduction

A speaker recognition system authenticates the identity of claimed users from a speech utterance (Mak and Chien, 2020). For a given speech segment called enrollment and a speech segment from a claimed user, the speaker recognition system will determine automatically whether both segments belong to the same speaker or not. The state-of-the-art speaker recognition systems mainly use Deep Neural Networks (DNN) to extract fixed-length speaker discriminant representations called speaker embeddings. The notable speaker embedding extractors that are used in the literature are TDNN (Snyder et al., 2018), ResNet (He et al., 2016), ECAPA-TDNN (Desplanques et al., 2020) and MFA-Conformer (Zhang et al., 2022).

The DNN-based speaker verification systems perform well in general, but there are some challenges that reduce their performance dramatically. Far-field speaker recognition is among the well-known challenges facing speaker recognition systems (Zheng et al., 2022). The far-field challenge is intertwined with other variabilities such as noise and reverberation leading to attenuated and distorted signal (Taherian et al., 2019).

Having well-designed benchmarks for challenging situations could reveal the weaknesses of speaker recognition systems and foster the research to address them. In this paper, we introduce a new single/multi-channel far-field speaker recognition benchmark called RoboVox. The Robovox benchmark is concerned with doing far-field speaker verification from speech signals recorded by a mobile robot at variable distances in the presence of noise and reverberation.

Although there are some benchmarks in this domain such as VoiCes (Richey et al., 2018) and FFSVC (Zheng et al., 2022), they have some deficiencies that our benchmark aims to address. A main drawback of the VoiCes is that it was recorded from replayed signals whereas our dataset is recorded with people speaking in real noisy conditions.

The FFSVC is another far-field speaker recognition benchmark that is recorded for smart home scenarios (Zheng et al., 2022). Dipco is another far-field speaker recognition benchmark that is derived from the Dipco corpus and replicates a scenario where a group of people are in an interactive conversation while having dinner in a home environment (Rouvier and Mohammadamini, 2022). Both FFSVC and Dipco are simulating a home environment, we extend these scenarios to workplace environments where having severe acoustical conditions is more probable.

In the robotics domain, there are other variabilities that have not been addressed in previous benchmarks: the robot's internal noise and the angle between the speaker and the robot. Furthermore, the speech signal has been recorded for different distances between the speaker and the robot. In the proposed challenge the following variabilities are present:

- **Ambient noise leading to low signal-to-noise ratios (SNR):** The speech signal is distorted with noise from fans, air conditioners, heaters, computers, etc.

LREC-COLING 2024, pages 14152–14156

- **Internal robot noises (robot activators):** The robot's activator noise reverberates on the audio sensors and degrades the SNR.
- **Reverberation**: The phenomena of reverberation due to the configuration of the places where the robot is located. The robot is used in different rooms with different surface textures and different room shapes and sizes.
- **Distance:** The distance between the robot and speakers is not fixed.
- **Babble noise:** The potential presence of several speakers speaking simultaneously.
- **Angle:** The angle between speakers and the robot's microphones

In the evaluation part of this paper, we will show that the mentioned variabilities facing the state-of-the-art speaker recognition systems reduce their performance dramatically and in comparison to the discussed scenarios in FFSVC and Dipco, the RoboVox is a more challenging benchmark. In the following, the dataset description comes in Section 2. In section 3 the evaluation protocols are discussed and section 4 and 5 describe the baseline system and the obtained results respectively.

2. Dataset description

Robovox is a French corpus recorded by a mobile robot (E4) in the framework of the ANR project RoboVox. The robot is equipped with a speaker recognition system in noisy environments. There are three microphones on the angles of the robot (Micro #1, Micro #2, Micro #3). The fourth microphone is embedded inside the robot (Micro #4). Another microphone serves as a ground truth microphone (Micro #5). The ground truth microphone is close to the mouth of the speaker. The microphones are depicted in Figure 1. The speech files are recorded from conversations between the robot and the speakers. The robot utilizes a loudspeaker positioned beneath the robot to articulate its utterances.
The dataset includes 78 speakers. The number of conversations between the robot and the speakers is between 24 and 36 which results in 2221 conversations. In each conversation, there are 5 dialogues (speaker turns) on average. Therefore, the total number of recorded dialogues is $\simeq 11,000$. The average length of each dialog is 3.6 seconds. Each recording has 8 channels. The channel information is as follows:

- **Channel 1 to 3:** The microphones on the angels of the robot;
- **Channel 4:** The microphone embedded inside the robot;
- **Channel 5:** The ground truth microphone which is close to the speaker;
- **Channel 6:** This is an unused channel;
- **Channel 7 and channel 8:** These channels include the robot's turns.

It is worth noting that having a clean signal recorded by Channel 5, enables us to have the best-expected baseline system and allows us to know the amount of performance degradation for far-field microphones.
The files are recorded from different distances in different acoustical environments with the main following settings:

- **1m, 2m and 3m:** The distance between the speaker and the robot in meters.
- **hall, open space, small room (open/close) and medium room (open/close):** The sessions are recorded in the different rooms/environments with the door open or closed in meeting rooms.
- **wall, center, and corner:** The robot placed close to a wall (or window), in the center of the room, or in the corner respectively. Severe reverberation can be spotted.
- **calm or noisy:** Level of noise in the environment.

The percentage of recorded conversations in terms of the distance, the environment, the location of the robot, and the noise is shown in Figure 2.

3. Evaluation Protocol

In this section, the details of protocols for both single-channel and multi-channel tracks are described.

3.1. Single-channel track

In the far-field single-channel, the best channel (i.e. channel 5) will be used for enrollment. For each speaker, three dialogues are used as enrollment. If a dialogue in a session is chosen as enrollment, the remaining dialogues in that session will not be used in the test. The microphones located on the angles of the robot and the embedded microphone inside the robot are evaluated separately.

3.2. Multi-channel track

In this track, the best channel (i.e. channel 5) is used for enrollment. All channels except the best channel (i.e. channel 5) are used for the test. The test and enrollment dialogues are the same as the single-channel track.

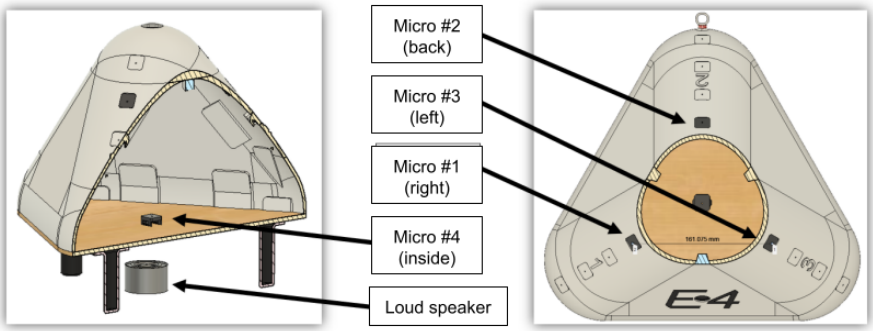

Figure 1: Robovox (E4): a mobile robot

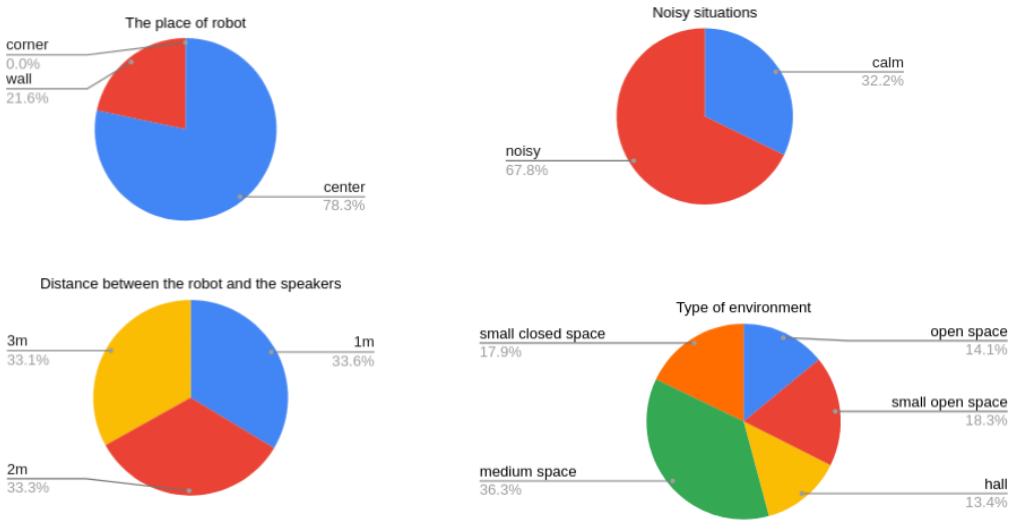

Figure 2: RoboVox dataset specifications

Firstly the dialogues extracted from the conversation between the robot and the speaker using the annotation provided for each conversation. First of all, the dialogues shorter than 2 seconds are discarded. Three dialogues longer than 4 seconds are chosen randomly per speaker as enrollment. The remaining dialogues from the chosen session for enrollment are filtered out in order to not be used in the test. The remaining part of the dialogues are used as a test. From these test and enrollment files, 2.47 million trials were generated. In order to have a 90/10% target/non-target, 30k target trials are merged with 270k non-targets. The details of the protocol are shown in Table 1.

Test	Enrollment	Trials	Target
10,332	225	300k	30k

Table 1: RoboVox evaluation protocol

4. Baseline system

The speaker embedding extractor used in this paper is a variant based on ResNet-34 (Rouvier and Bousquet, 2021). In the first and second blocks, we used the squeeze and excitation mechanism. The loss function is the angular additive margin with a margin equal to 0.4. The size of the feature maps are 32, 64, 128, and 256 for the 4 ResNet blocks (Table 2). We use stochastic gradient descent with a momentum equal to 0.9, a weight decay equal to 2.10^{-4}, and an initial learning rate equal to 0.2. The batch size was set to 128.

The speaker embedding extractor is trained on the development partition of the VoxCeleb2 dataset which contains speech utterances from 5,994 speakers (Nagrani et al., 2020). The extractor is trained with 4-second chunks of training samples and their augmented version with noise and reverberation as described in (Snyder et al., 2015). As input, we used 60-dimensional filter-banks. In order to remove silence, a simple energy-based VAD is applied to the acoustic features (Povey

Layer name	Structure	Output
Input	–	$60 \times 400 \times 1$
Conv2D-1	3×3, Stride 1	$60 \times 400 \times 32$
SE-Block-1	$\begin{bmatrix} 3 \times 3, 32 \\ 3 \times 3, 32 \end{bmatrix} \times 3$, Stride 1	$60 \times 400 \times 32$
SE-Block-2	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 4$, Stride 2	$30 \times 200 \times 64$
ResNetBlock-3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 6$, Stride 2	$15 \times 100 \times 128$
ResNetBlock-4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 3$, Stride 2	$8 \times 50 \times 256$
Pooling	–	8×256
Flatten	–	2048
Dense1	–	256
Dense2 (Softmax)	–	N
Total	–	–

Table 2: The baseline ResNet-34 architecture.

et al., 2011).

We report results in terms of Equal Error Rate (EER) and the detection cost function (DCT)(Sadjadi et al., 2021). The DCT is defined as Equation 4:

$$C_{DET}(\theta) = C_{Miss} \times P_{Miss|Target}(\theta) \times P_{Target} + C_{FA} \times P_{FA|NonTarget}(\theta) \times (1 - P_{Target}) \quad (1)$$

where: θ is a decision threshold, C_{Miss} is the cost of false rejection, C_{FA} is the cost of false acceptance, P_{Target} is the prior probability of target speakers. The parameters for DCT are defined in Table 3. The first scenario is for using the robot during the day when the probability of target speakers is high and the second scenario is for using it during the night with a low probability for target speakers and high C_{FA}. The average of the two scenarios will be the final DCT score.

$C_{DET}(\theta)$	P_{Target}	C_{Miss}	C_{FA}
1	0.8	1	20
2	0.01	10	100

Table 3: C_{DET} parameters

5. Results

Firstly the trained speaker embedding extractor is evaluated on standard Voxceleb-E cleaned protocol. In this case, we achieved 1.12 in terms of EER. These results show the efficiency of the baseline system. In Table 4, the obtained results for both single-channel and multi-channel tracks are reported. For channel 5 which is the best channel the EER is equal to 9.29. The main possible reason for these results can be the short duration of the test files. In channel 4, which is the worst case the EER is equal to 18.22. The same behavior is observed for the rest of the channels. If we compare the results with the baseline results reported for DipCo (Rouvier and Mohammadamini, 2022) and FFSVC (Zheng et al., 2022) corpus, we can see that RoboVox is a more challenging situation which makes it a rigorous benchmark for the evaluation of speaker embedding extractors. For example, the EER in the Dipco single-channel track is 5.84 while in the RoboVox is 18.22. In the multi-channel case, the EER and DCT are calculated based on the average cosine score of four far channels (i.e. Channel 1-4).

Channel	EER	DCT
Channel 1	15.79	0.92
Channel 2	15.63	0.87
Channel 3	15.74	0.88
Channel 4	18.22	0.91
Channel 5	9.29	0.73
Multi-channel	15.06	0.86

Table 4: EER and DCT

6. Copyrights

This audio database is made available under the terms of the Creative Commons Attribution NonCommercial-ShareAlike 4.0 International License [1]. The dataset will be available for researchers by asking the providers from this link [2].

7. Acknowledgements

This research was made possible by the financial support provided by Agence Nationale de la Recherche (ANR) in the framework of the RoboVox project.

8. Conclusion

In this paper, we introduced a new benchmark for far-field single-channel and multi-channel speaker recognition for a Mobile Robot. We did an evaluation for both single-channel and multi-channel cases leveraging a competitive state-of-the-art ResNet speaker embedding extractor. The significant performance reduction of the speaker recognition systems in real far-filed applications could foster the researchers to address the far-field recognition in the presence of noise and reverberation.

9. Bibliographical References

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834.

[1] https://creativecommons.org/licenses/by-nc-sa/4.0/

[2] https://huggingface.co/datasets/aranemini/RoboVox

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Man-Wai Mak and Jen-Tzung Chien. 2020. *Deep Learning Models*, page 115–168. Cambridge University Press.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech Language*, 60:101027.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Colleen Richey, Maria A. Barrios, Zeb Armstrong, Chris Bartels, Horacio Franco, Martin Graciarena, Aaron Lawson, Mahesh Kumar Nandwana, Allen Stauffer, Julien van Hout, Paul Gamble, Jeffrey Hetherly, Cory Stephenson, and Karl Ni. 2018. Voices Obscured in Complex Environmental Settings (VOiCES) Corpus. In *Proc. Interspeech 2018*, pages 1566–1570.

Mickael Rouvier and Pierre-Michel Bousquet. 2021. Studying squeeze-and-excitation used in cnn for speaker verification. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1110–1115.

Mickael Rouvier and Mohammad Mohammadamini. 2022. Far-field speaker recognition benchmark derived from the DiPCo corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1955–1959, Marseille, France. European Language Resources Association.

Omid Sadjadi, Craig Greenberg, Elliot Singer, Lisa Mason, and Douglas Reynolds. 2021. Nist 2021 speaker recognition evaluation plan.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. ArXiv:1510.08484v1.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Hassan Taherian, Zhong-Qiu Wang, and DeLiang Wang. 2019. Deep Learning Based Multi-Channel Speaker Recognition in Noisy and Reverberant Environments. In *Proc. Interspeech 2019*, pages 4070–4074.

Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung yi Lee, and Helen Meng. 2022. MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification. In *Proc. Interspeech 2022*, pages 306–310.

Yu Zheng, Jinghan Peng, Yihao Chen, Yajun Zhang, Jialong Wang, Min Liu, and Minqiang Xu. 2022. The SpeakIn Speaker Verification System for Far-Field Speaker Verification Challenge 2022. In *Proc. The 2022 Far-field Speaker Verification Challenge (FFSVC2022)*, pages 15–19.