ROBOVOX: FAR-FIELD SPEAKER RECOGNITION BY A MOBILE ROBOT (EVALUATION PLAN)

Mohammad Mohammadamini[1], *Mickael Rouvier*[1], *Driss Matrouf*[1], *Jean-François Bonastre*[1]
Romain Serizel[2] , Denis Jouvet[2], Théophile Gonos[3]

[1]Avignon University, LIA (Laboratoire Informatique d'Avignon), Avignon, France
[2] University of Lorraine, CNRS, Inria, Loria, Nancy, France,
[3] A.I.Mergence, Paris, France,

1. INTRODUCTION

A speaker recognition system authenticates the identity of claimed users from a speech utterance. 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 decision to accept or reject a speaker will be made by comparing speaker embeddings.

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. The far-field challenge is intertwined with other variabilities such as noise and reverberation. Two main categories of speaker recognition systems are text-dependent speaker recognition and text-independent speaker recognition. In a text-dependent speaker recognition system, the speaker's voice is recorded from predefined phrases, while, in text-independent speaker recognition, there is no constraint on the content of the spoken dialogue. The task of the IEEE Signal Processing Cup 2024 is text-independent far-filed speaker recognition under noise and reverberation for a mobile robot.

2. TASK DESCRIPTION

The Robovox challenge 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 and FFSVC, they don't cover variabilities in the domain of robotics such as the robot's internal noise and the angle between the speaker and the robot. The VoiCes dataset is replayed speech recorded under different acoustical noises. A main drawback of the VoiCes is that it was recorded from played signals whereas our dataset is recorded with people speaking in noisy environments. The FFSVC is another far-field speaker recognition benchmark. However, these benchmarks helped the community significantly, we are introducing a new benchmark for far-field speaker recognition systems in order to address some new aspects. Firstly, our goal is to perform speaker recognition in a real application for the domain of mobile robots. In this 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.
- **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 and it is possible for the robot to move during the recognition.
- **Babble noise:** The potential presence of several speakers speaking simultaneously.
- **Angle:** The angle between speakers and the robot's microphones

In this challenge, two tracks will be proposed:

- **Far-field single-channel tracks**: In this task, one channel is used to perform the speaker verification.

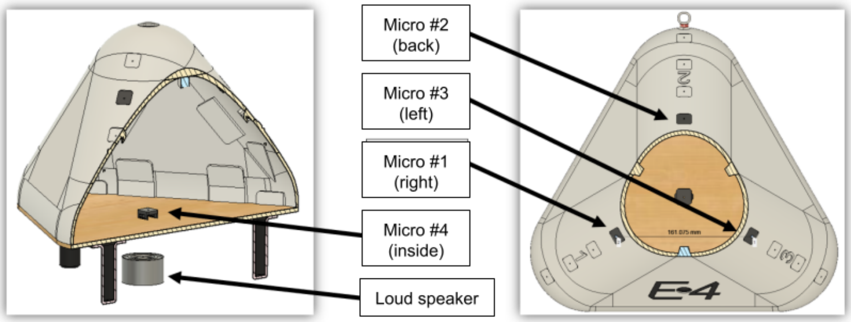

Fig. 1. Robovox (E4): a mobile robot

The main objective is to propose novel robust speaker recognition pipelines to tackle the problem of far-field speaker recognition in the presence of reverberation and noise.

- **Far-field multi-channel tracks**: In this task, several channels are used to perform speaker verification. The main objective is to develop algorithms that improve the performance of multi-channel speaker verification systems under severe noise and reverberation.

3. DATASET DESCRIPTION

In this challenge, we introduce a novel benchmark that complements previous works and aims at fostering research in **far-field single-channel** and **multi-channel** speaker verification. We will propose an evaluation benchmark in which the voice dialogues are recorded by a robot in various acoustic conditions.

The 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 is used as a ground truth microphone (Micro #5). The ground truth microphone is close to the mouth of the speaker. The microphones are depicted in Fig 1. The speech files are recorded from conversations between Robovox and speakers. Robovox 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 2219 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:** microphones on the angels of the robot;
- **Channel 4:** microphone embedded inside the robot;
- **Channel 5:** ground truth microphone which is close to the speaker;
- **Channel 6:** Unused channel;
- **Channel 7 and channel 8:** Are robots dialogues 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. An example of a recorded signal spectrum is depicted in Figure 2:

The files are recorded from different distances in different acoustical environments with the main following settings:

- **1m, 2m and 3m:** Distance of the speaker from the robot: respectively 1, 2, and 3 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 close in meeting rooms.
- **wall, center, and corner:** The robot is 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.

3.1. Licence

This audio database is made available under the terms of the Creative Commons Attribution NonCommercial-ShareAlike 4.0 International License. This means that you are free to share (copy, distribute, and transmit the work) and remix (adapt the work), as long as you credit the original authors,

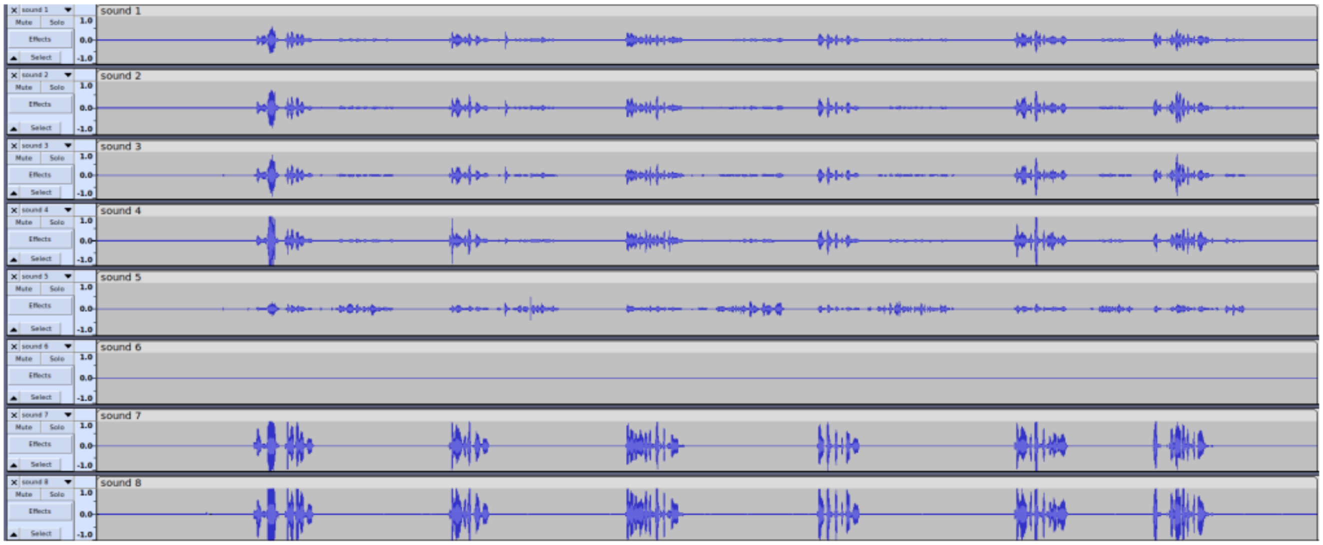

Fig. 2. An example of a recorded speech signal by Robovox

do not use this work for commercial purposes, and share any derivative works under a similar license. For more information, please visit the Creative Commons website.

3.2. Evaluaition protocols

3.2.1. Task1: Far-field single-channel tracks

In the Far-field single-channel track, two channels will be used. The best channel (i.e. channel 5) will be used for enrollment. For each speaker, three dialogues are used as enrollment. If a session is chosen as enrollment the remained dialogues in that session will not be used in the test. The remaining part of the dialogues are used as the test. The most challenging channel (i.e. channel 4) is used for the test.

3.2.2. Task 2: Far-field multi-channel tracks

In this task, all channels can be used for enrollment. Channel 5 is not allowed to be used as the test. Similar to the first task a few dialogues are used in enrollment and the rest of the dialogues (Except dialogues from sessions used in the enrollment) will be used as the test.

3.3. Data organization

The data is organized as follows:

main directory/
readme.txt
data/
samples/
single-channel/
enrollment/
test/
single-channel-trials.trl
multi-channel/
enrollment/
test/
multi-channel-trials.trl
docs/
readme
robovox-challenge.pdf

In the *samples* directory you can find speech sample files from two speakers. You are allowed to use the sample files for training your model or use their information for other tasks such as data simulation or data augmentation. For both single-channel and multi-channel tasks the preprocessing is done on the recorded sessions. The *data/single-channel/test* directory contains the noisy dialogues extracted from channel 4 which is the worst case. The *data/single-channel/enrollment* contains enrollment files for single-channel tasks coming from the ground-truth microphone which has the highest quality among other channels. The *data/multi-channel/enrollment* and *data/multi-channel/test* include the enrollment and test files for multi-channel tasks respectively. The single-channel trials will be found in *single-channel/single-channel-trials.trl* and multi-channel trials are in *multi-channel/multi-channel-trials.trl* file.

In order to get the dataset the participants should register on Codebench and get the data from the files tab.

3.4. Evaluation metrics

The decision cost function (DCF) is used as a primary evaluation metric which is defined as:

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

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. We consider two sets of parameters to calculate $C_{DET}(\theta)$. The parameters are listed in Table 1. The first line shows the parameters using the robot during the day when the probability of having a target speaker is high and the cost of accepting a non-authorized person is less in comparison to night. The last row shows the parameters for using the robot during the night with a low P_{Target} and high C_{FA}.

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

Table 1. C_{DET} parameters

The final score will be calculated as the average of C_{DET} for two sets of parameters listed in Table 1.

The Equal Error Rate (EER) will be used as a secondary evaluation metric that is defined as an operating point θ, where $FAR = FRR$. False Acceptance Rate (FAR) is the number of False Acceptance (FA) errors divided by the total number of nontarget trials(imposters). And, the False Rejection Rate (FRR) is the number of False Rejection (FR) errors divided by the total number of legitimate trials.

3.5. Participants submissions

3.5.1. Results

We provided a scoring and ranking tool on Codebench. The scores will be calculated automatically. The participants should make a valid submission for the single-channel task. A valid submission should contain a score for all trials. A submission with missed trials will not be considered.

Put the scores in a single text file with this name: yourID.txt. Then put it in a compressed *.zip file and submit it from the submission tab.

The format of a submission is as follows:

enroll_ID <TAB>*test_ID* <TAB>*score*

where *enroll_ID* is the id of an enrollment file, *test_ID* is the id of a test file, and *score* is the cosine distance between test and enrollment files.

A sample of the result file is shown here:

```
spk_72   93dfe359e7   0.6813565
spk_12   4426dfd7fe   0.7223774
spk_23   bb6235bfce   0.9346804
```

Note: The multi-channel track will not be considered for ranking the participants. We are proposing this track to foster the research in multi-channel speaker recognition.

3.5.2. Technical report

Each group should submit a technical report that includes:

- **System description:** This part should describe the front-end such as hand-made or DNN-based features, the VAD, the speaker embedding extraction architecture, etc.
- **Sample files:** Along with the evaluation protocols recorded sessions from two speakers are provided as sample files. The participants can use these files to adapt their system or use the available information for data augmentation etc. In the case of using these files report the way you exploited them.

The technical report is obligatory for introducing the winners. Please send your report before final submission date (5 February 2024) in a PDF format to the provided email of the contact person with this subject: yourCodebenchID_SPCUP2024_report

4. GENERAL DETAILS AND RULES ABOUT THE COMPETITION

The Signal Processing Cup (SP Cup) competition is held annually and encourages teams of students to work together to solve real-world problems using signal processing methods and techniques. Each year, three final teams are chosen to present their work during ICASSP to compete for the US$5,000 grand prize!

4.1. Team Formation and Eligibility

Each team participating should be composed of one faculty member or someone with a Ph.D. degree employed by the university (the Supervisor), at most one graduate student (the Tutor), and at least three, but no more than ten undergraduate students. At least three of the undergraduate team members must hold either regular or student memberships of the IEEE Signal Processing Society. Undergraduate students who are in the first two years of their college studies, as well as high school students who are capable of contributing, are welcome to participate in a team. A participant cannot be on more than one team.

4.2. Prize for Finalists

The three teams with the highest performance in the open competition will be selected as finalists and invited to participate in the final competition at ICASSP 2024. The champion team will receive a grand prize of $5,000. The first and the second runner-up will receive a prize of $2,500 and $1,500, respectively, in addition to travel grants and complimentary conference registrations.

Up to three student members from each finalist team will be provided travel support to attend the conference in person. In-person attendance of the physical conference is required for reimbursement. Complimentary conference registration for the three finalist team members from each team who present at ICASSP. These complimentary conference registrations cannot be used to cover any papers accepted by the conference. If you are one of the three finalist team members from each team and wish to receive complimentary registration and/or conference banquet access, you must email Jaqueline Rash, Jaqueline.rash@ieee.org, with this information once your team has been selected as a finalist. The three finalist team members from each team will also be invited to join the Conference Banquet and the SPS Student Job Fair, so that they can meet and talk to SPS leaders and global experts. Please note registration to the Conference Banquet and Student Job Fair is limited and based on availability.

4.3. Sponsers

We gratefully acknowledge MathWorks, Inc. for their continued support of IEEE SP Cup. Participating students are encouraged to download the complimentary Mathworks Student Competitions Software for use in the competition.

5. IMPORTANT DATES

- Challenge announcement 1 November 2023
- Final submission due: 5 February 2024
- Finalists announcement: 14 February 2024
- Final competition at ICASSP 2024: April 14-19, 2024

6. CONTACT PERSON

Mohammad MOHAMMADAMINI
Email: mohammad.mohammadamini@univ-avignon.fr
Tel: +33 7 80 86 21 64