Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1955–1959
Marseille, 20-25 June 2022

Far-Field Speaker Recognition Benchmark Derived From The DiPCo Corpus

Mickael Rouvier, Mohammad Mohammadamini
LIA - Avignon University
339 Chemin des Meinajaries, 84000, Avignon
{mickael.rouvier, mohammad.mohammadamini}@univ-avignon.fr

Abstract

In this paper, we present a far-field speaker recognition benchmark derived from the publicly-available DiPCo corpus. This corpus comprises three different tasks that involve enrollment and test conditions with single- and/or multi-channel recordings. The main goal of this corpus is to foster research in far-field and multi-channel text-independent speaker recognition. Also, it can be used for other tasks such as dereverberation, denoising, and speech enhancement. In addition, we release recipes implemented with Kaldi and SpeechBrain libraries to facilitate further research. We validate the evaluation design with a single-microphone state-of-the-art speaker recognition system (i.e. ResNet-101). The results show that the proposed tasks are very challenging. And we hope these resources will inspire the speech community to develop new methods and systems for this challenging domain.

Keywords: speaker recognition, benchmark, far-field, DiPCo corpus

1. Introduction

Speaker Recognition refers to the authentication of the claimed users from their voice (Bimbot et al., 2004). Speaker recognition systems have been used in several applications such as speaker diarization (Rouvier and Meignier, 2012), forensics (Campbell et al., 2009) or voice dubbing (Gresse et al., 2017). Although the state-of-the-art DNN-based systems perform well in adverse environments, still 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 challenges because the far distance signal will be impacted more by other distortions such as noise and reverberation. Having a real-world evaluation benchmark for far-field speaker recognition is highly demanded by the speech community. Over the years, a few far-field corpus or far-field challenges have emerged.

In (Qin et al., 2020), a Far-Field Speaker Verification Challenge (FFSVC) was proposed with a special focus on far-field distributed microphone arrays under noisy conditions in real scenes. In FFSVC three tasks are proposed: far-field text-dependent speaker verification from a single microphone array, far-field text-independent speaker verification from a single microphone array, and far-field text-dependent speaker verification from distributed microphone arrays.

The VOiCES challenge (Nandwana et al., 2019) focuses on the robustness of the replayed speech in the presence of reverberation and background noises. VOiCES corpus was collected by recording played audio from high-quality loudspeakers in real rooms, capturing natural reverberation. The retransmitted corpus is LibriSpeech which is used as the clean speech source, while television, music, or babble played simultaneously from another loudspeaker as background noise. The main deficiency of VOiCES is that it doesn't match real-world scenarios.

In (Garcia-Romero et al., 2019), the authors propose a speaker recognition benchmark derived from the CHIME-5 challenge. The benchmark comprises four tasks for enrollment and test conditions with single-speaker and/or multi-speaker recordings. Additionally, it supports performance comparisons between close-talking vs. distant/far-field microphone recordings and single-microphone vs. microphone-array approaches. Unfortunately, despite its attractiveness, it has never been released to the public.

In this paper, we present a novel benchmark that complements previous works and aims at fostering research in multi-channel speaker recognition. The speaker recognition benchmark is derived from the publicly available DiPCo corpus (Van Segbroeck et al., 2019), which was initially designed to foster research in the field of noise-robust and distant microphones for automatic speech recognition. Here we reuse these data to build a speaker evaluation dataset to explore the performance of speaker verification systems for conversational speech in noisy environments and multi-microphone distant/far-field. In addition, we release Kaldi and SpeechBrain [1] recipes to facilitate further research.

The paper is organized as follows. Section 2 present the DiPCo corpus and the far-field speaker verification benchmark derived from the DiPCo corpus. We present in Section 3 the three different tasks. Experiments and results are presented in Section 4 before concluding in Section 5.

[1] https://dipco-sre.univ-avignon.fr/

2. Dataset

2.1. Summary of the DiPCo corpus

The Dinner Party Corpus (DiPCo) is a speech database that replicates the scenario where a group of people is having an interactive conversation while having dinner in a simulated home environment. The corpus consists of multiple sessions recorded in the same room over multiple days and with different groups of participants. More precisely, the corpus contains the collection of 10 sessions in which 4 persons have a natural conversation over dinner and has a total of 32 unique speakers.
At the beginning of each session, participants were getting food at the buffet and then moved to the dining table. All participants are seated around a dining table, and in each session, music playback started at a given time mark.

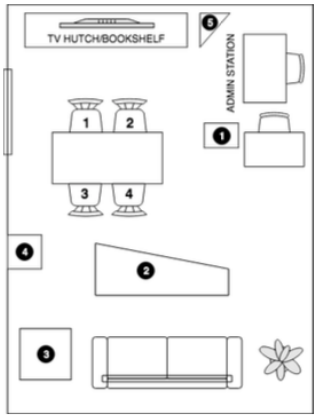

Figure 1: Layout of the room in which the sessions were recorded (figure source : (Van Segbroeck et al., 2019)).

In this corpus, all participants have been simultaneously recorded with a single-channel close-talking (1 close-talking) and far-field microphone devices (5 far-field microphones). Figure 1 shows the floor plan and layout of the room and Table 1 gives the distance measure between the participants and the microphone array devices (in mm). The far-field devices were equipped with a microphone array consisting of 7 microphone channels. The 7-microphone array was configured as illustrated in Figure 2. The 6 microphones were uniformly placed on the perimeter of a circle of radius 35 millimeters, the 7th microphone was placed at the center of the circle.
All audio was distributed at a 16 KHZ sampling rate. The close-talk recordings of all speakers were manually transcribed and sentence boundaries were provided. The microphone recordings per session were all time-synchronized.

2.2. Derived far-field speaker corpus

To create the derived far-field speaker corpus, we used the 10 sessions assigned to the dev and test partition of the DiPCo corpus.

Persons	Device				
	1	2	3	4	5
1	1600	2240	3825	2900	1760
2	1990	2130	3950	3100	1760
3	1820	1520	2900	2030	2780
4	1300	1120	3100	2520	2820

Table 1: Distance between the participants and the microphone array devices (in mm).

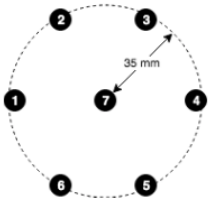

Figure 2: Configuration of the 7-microphone array (figure source : (Van Segbroeck et al., 2019)).

The extraction of enrollment and test segments is based on the segmentation given in the DiPCo corpus. In the DiPCo corpus, the segment is defined as a continuous audio chunk where the utterance segment boundaries are determined by a transcriber that searches a logical timestamp, such as the beginning of a new sentence. Leveraging this segmentation, we removed all overlapping speech regions so that each segment contains the voice of only one speaker. And finally, we removed all the segments with less than 3 seconds of voice and those that exceed 8 seconds of voice. Because doing speaker recognition for very short utterances becomes very challenging and it is not easy to differentiate the effect of noise and duration. In the same manner, for very long utterances even for far-field utterances, there is enough information for recognition.
For enrollment, we only use the segment audio from the close-talking recordings. And for the test, we use the different segment audios from the far-field microphone. Figure 3 shows the total duration of the enrollment utterances selected for each of the 31 speakers.
We note that we dismissed participant P13 because the close-talk microphone recording is noisier than others due to a microphone issue in the DiPCo corpus.
In order to create a more challenging far-field speaker recognition benchmark, we selected the pairs with the same gender (male and female) and the same nativeness (native and non-native) in trials. In other words, there are no cross-gender trials and no cross-nativeness trials.
In addition to this far-field speaker recognition benchmark, metadata is available to enable more analysis of the performance. The types of metadata include gender, the distance between the participants and the microphone array devices, and the recording environment (noise, laughter, music).
The corpus and metadata are available through

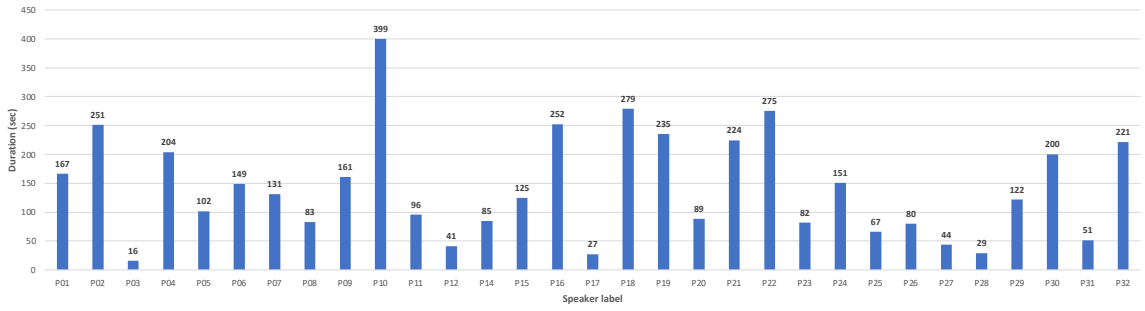

Figure 3: Total duration of the enrollment utterances selected for each of the 31 speakers in the corpus.

the website: `https://dipco-sre.univ-avignon.fr/`.

3. Tasks description

Our goal in this paper is to investigate the effect of single-channel and multi-channel speaker recognition in the context of far-field. We define three tasks based on three different conditions.

We note that the enrollment condition is the same for all three different tasks. The enrollment segments are extracted from the close-talking microphone. One speech segment is used to build the speaker model. In the test condition, the number of speech segments (and the source of the microphone) is different for each task. The test segments are extracted from the far-field microphone.

The three tasks from the far-field speaker benchmark are the following:

- **Task-1 - Far-field speaker recognition from single microphone**: The task corresponds to the single-channel test segment. One speech segment is used to build the model of the target speaker. The test segments are extracted from one of five far-field microphones and at each time the 7th channel is selected (the one which is placed at the center of the circle).
- **Task-2 - Far-field speaker recognition from single microphone array**: This task corresponds to the multi-channel test segment. Seven speech segments are used to build the model of the target speaker. The test segments are extracted from one of five far-field microphones and we selected all the channels.
- **Task-3 - Far-field speaker recognition from distributed microphone**: This task corresponds to the multi-microphone test segment. Four speech segments are used to build the model of the target speaker. The test segment is extracted from four of five far-field microphones and at each time the 7th channel is selected.

Table 2 provides some information about the trial conditions for the three tasks, including the target and impostor trial counts.

4. Baseline system

This section provides baseline results for the three different tasks of the far-field speaker recognition benchmark derived from the DiPCo corpus.

4.1. Training and Evaluation datasets

The x-vector extractors are trained on the VoxCeleb2 dataset (Chung et al., 2018). Only the development partition of VoxCeleb2, which contains speech excerpts from 5,994 speakers with a 16 KHz sampling rate, is used.

Firstly, the trained x-vectors are assessed on the Speakers in the Wild (SITW) core-core task (McLaren et al., 2016) and Voxceleb1-E Cleaned with a 16 KHZ sampling rate. We show these results to provide an indication of the strength of the baseline system. Note that the development set of VoxCeleb2 is completely disjoint from the VoxCeleb1 dataset (*i.e.* no common speakers).

Finally, the trained x-vectors are assessed on the three different tasks of the far-field speaker benchmark derived from the DiPCo corpus.

We report results in terms of Equal Error Rate (EER) and the minimum detection cost function (minDCF) with $C_{miss} = 1$, $C_{fa} = 1$ and $P_{target} = 0.01$.

4.2. Implementation details of x-vector extractor

The x-vector extractor used in this paper is a variant based on ResNet-101. The architecture of the x-vector extractor is described in Table 3. 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., 2018) which is available as a part of a Kaldi recipe. As input, we used 60-dimensional filter-banks. The x-vectors are 256-dimensional and the loss function is angular additive margin with a scale equal to 30 and margin equal to 0.4. The size of the feature maps are 128, 128, 256, and 256 for the 4 ResNet blocks. 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, training on 4 GPUs in parallel. The implementation is based on PyTorch and the model training takes about 2 days. In order to

	# Pairs	#Positive Pairs	# Utterances	Segment length (s)
Task-1	900,000	150,000	4,405	3.51/5.03/7.99
Task-2	900,000	150,000	4,405	3.51/5.03/7.99
Task-3	900,000	150,000	30,835	3.51/5.03/7.99

Table 2: Statistics of the far-field speaker recognition(Task 1-3). # Pairs refers to the number of evaluation trials pairs, **# Positive Pairs** refers to the number of evaluation trials positive, **# Utterances** refers to the total number of unique speech segment in the test set and **# Segment length** are reported as min/mean/max.

remove silence and low energy speech segments, a simple energy-based VAD is used based on the C0 component of the acoustic feature.

Layer name	Structure		Output
Input	–		$60 \times 400 \times 1$
Conv2D	3×3, Stride 1		$60 \times 400 \times 128$
ResBlock-1	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \\ 3 \times 3, 128 \\ SE, [128, 60] \end{bmatrix}$	$\times 3$, Stride 1	$60 \times 400 \times 128$
ResBlock-2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \\ 3 \times 3, 128 \\ SE, [128, 30] \end{bmatrix}$	$\times 4$, Stride 2	$30 \times 200 \times 128$
ResBlock-3	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix}$	$\times 23$, Stride 2	$15 \times 100 \times 256$
ResBlock-4	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix}$	$\times 3$, Stride 2	$8 \times 50 \times 256$
Pooling	–		8×256
Flatten	–		2048
Dense1	–		256
Dense2 (Softmax)	–		N
Total	–		–

Table 3: The ResNet-101 architecture. In the last row, N is the number of speakers. Batch-norm and ReLU layers are not shown. The dimensions are (Frequency×Channels×Time). The input is comprised of 60 filter banks from speech segments. During training we use a fixed segment length of 400 frames.

4.3. Results

Table 4 shows the performance obtained by the baseline system on VoxCeleb1-E Cleaned and SITW. These datasets are extensively used in speaker recognition area. We show these results to provide an indication on the strength of the baseline system. We observe that the system obtain on VoxCeleb1-E Cleand and SITW an EER close to 1% and a DCF close to 0.1.

	EER(%)	DCF
VoxCeleb1-E	1.02	0.115
SITW	1.15	0.100

Table 4: Performance obtained by the baseline system on Voxceleb1-E Cleaned and SITW.

Table 5 shows the performance on the derived DiPCo corpus across the three different tasks. We observe that the baseline system obtained on the Task-1, Task-2 and Task-3 an EER respectively of 5.84%, 4.89% and 3.65%. The EER is more important for the Task-1 because the test is constituted of one speech segment. We can see that this corpus is extremely challenging. The EER is 3 to 6 times greater than VoxCeleb1-E Cleaned or SITW.

	EER(%)	DCF
Task-1	5.84	0.369
Task-2	4.89	0.308
Task-3	3.65	0.263

Table 5: Performance obtained by the baseline system across the three different tasks of the derived DiPCo corpus.

5. Conclusions

This article provides details on the derived DiPCo far-field speaker recognition benchmark. This corpus is publicly available and is designed to foster robustness against the artifacts introduced by far-field and multi-channel recordings. Also, this corpus can be readily used for dereverberation, denoising and speech enhancement. We release a Kaldi and SpeechBrain system to facilitate further research. The performance obtained by the state-of-the art system shows that the different proposed tasks are extremely challenging. We hence encourage the research community to develop new methods and systems for this challenging domain.

6. Acknowledgement

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011013257 made by GENCI and was supported by the ANR agency (Agence Nationale de la Recherche), RoboVox project (ANR-18-CE33-0014)

7. Bibliographical References

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., and Reynolds, D. A. (2004). A tutorial on text-independent speaker verification. *EURASIP Journal on Advances in Signal Processing*, 2004(4):101962.

Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., and Matrouf, D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, 26(2):95–103.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition.

Garcia-Romero, D., Snyder, D., Watanabe, S., Sell, G., McCree, A., Povey, D., and Khudanpur, S. (2019). Speaker recognition benchmark using the chime-5 corpus. In *Interspeech*, pages 1506–1510.

Gresse, A., Rouvier, M., Dufour, R., Labatut, V., and Bonastre, J.-F. (2017). Acoustic pairing of original and dubbed voices in the context of video game localization. In *Interspeech*.

McLaren, M., Ferrer, L., Castan, D., and Lawson, A. (2016). The speakers in the wild (sitw) speaker recognition database. In *Interspeech*, pages 818–822.

Nandwana, M. K., Van Hout, J., Richey, C., McLaren, M., Barrios, M. A., and Lawson, A. (2019). The voices from a distance challenge 2019. In *Interspeech*, pages 2438–2442.

Qin, X., Li, M., Bu, H., Rao, W., Das, R. K., Narayanan, S., and Li, H. (2020). The interspeech 2020 far-field speaker verification challenge. *arXiv preprint arXiv:2005.08046*.

Rouvier, M. and Meignier, S. (2012). A global optimization framework for speaker diarization. In *IEEE Odyssey - The Speaker and Language Recognition Workshop*.

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

Van Segbroeck, M., Zaid, A., Kutsenko, K., Huerta, C., Nguyen, T., Luo, X., Hoffmeister, B., Trmal, J., Omologo, M., and Maas, R. (2019). Dipco–dinner party corpus. *arXiv preprint arXiv:1909.13447*.