INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speaker Recognition Benchmark using the CHiME-5 Corpus

Daniel Garcia-Romero, David Snyder, Shinji Watanabe, Gregory Sell, Alan McCree, Daniel Povey, Sanjeev Khudanpur

Human Language Technology Center of Excellence & Center for Language and Speech Processing
The Johns Hopkins University, Baltimore, MD 21218, USA

dgromero@jhu.edu

Abstract

In this paper, we introduce a speaker recognition benchmark derived from the publicly-available CHiME-5 corpus. Our goal is to foster research that tackles the challenging artifacts introduced by far-field multi-speaker recordings of naturally occurring spoken interactions. The benchmark comprises four tasks that involve 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. We validate the evaluation design with a single-microphone state-of-the-art DNN speaker recognition and diarization system (that we are making publicly available). The results show that the proposed tasks are very challenging, and can be used to quantify the performance gap due to the degradations present in far-field multi-speaker recordings.

Index Terms: speaker recognition, multi-speaker, far-field speech, robustness.

1. Introduction

The speaker recognition community has greatly benefited from the evaluations hosted by the National Institute of Standards and Technology (NIST) since 1996. The data and benchmarks associated with these evaluations has facilitated fair system comparisons as well as monitoring the state-of-the-art performance over time. As a way to encourage participation, some of these databases are only shared freely to participants. Which has limited the accessibility of these resources to the larger research community. In the recent years, a new trend has emerged in which research groups are collecting and freely sharing corpora and evaluation protocols that focus on research aspects that complement the directions explored in NIST evaluations (mostly focused on telephone conversational speech). A pioneer of this trend was the Speakers in the Wild (SITW) [1] corpus and benchmark [2] organized by SRI in 2016. The corpus contains hand-annotated speech samples from open-source media acquired in unconstrained "wild" conditions. The benchmark involved single-speaker and multi-speaker recordings for both enrollment and test. Moreover, SITW allows using systems that operate at 16 KHz sampling rate.

Another recent development was the release of the largest publicly-available dataset combining the VoxCeleb-1 [3] and VoxCeleb-2 [4] corpora. The goal was to provide a labeled set (of similar size to those used by the face recognition community) to train Deep Neural Networks (DNNs). As a result, wideband speech (16 KHz) from more than 7000 speakers is now freely available. The speech was collected in similar conditions to the SITW corpus. The authors also provide an evaluation protocol [4] that focuses on a large number of test speakers (around 1000 compared to the 300 speakers in SITW). In this work, we use the VoxCeleb data to train our baseline system.

Additionally, the DeepMine database [5], mostly focused in Persian, consists of more than 1800 speakers with very good coverage of accents, speaker ages, and diverse mobile telephony. A subset of the speakers also spoke English, which allows cross-lingual studies. Along this line, the JSpeech [6] multi-lingual corpus expands this area by providing 1332 hours of chat group conversational speech in 47 languages from 12140 speakers.

Finally, the currently ongoing VOiCES [7] and Fearless Steps [8] challenges also explore interesting areas. VOiCES focuses on robustness to reverberation and background noises of replayed speech, while Fearless Steps defines a more holistic set of challenges using data from the Apollo-11 mission.

In this work, we contribute another dataset/benchmark that complements this rich ecosystem. In addition, we are releasing a Kaldi baseline system to facilitate further research. The benchmark is derived from the publicly available CHiME-5 corpus [9], which was initially designed to foster research in multi-microphone distant/far-field automatic speech recognition of multi-speaker, overlapping, conversational speech in noisy environments. Here we reuse this data to build a speaker recognition evaluation that explores those same conditions. In particular, the interesting qualities of this benchmark are that: i) it allows exploration of microphone-array beamforming/enhancement, and comparison against single-microphone approaches; ii) it is possible to compare the performance gap between a close-talking recording vs simultaneously recorded distant microphone versions; iii) it provides a great opportunity for speaker diarization techniques that can process multi-speaker enrollment and test segments.

2. Dataset

2.1. Summary of the CHiME-5 corpus

The CHiME-5 corpus [9] comprises twenty separate dinner parties taking place in real homes. Each gathering has four participants that are friends with each other and act naturally. Most of the participants attended two gatherings. The parties were organized in three phases which correspond to different locations and activities around the house: i) kitchen, meal preparation; ii) dining-room, eating the meal; iii) living-room, post-dinner conversation. Participants were allowed to move naturally and converse about topics of their choosing.

Each party was recorded using six Microsoft Kinect devices (denoted as U01, U02, U03, U04, U05, and U06) that were placed around the house to provide good coverage of the activities. Each device contains a 4-microphone array that was used to extract 4 channels of audio. In addition, each participant wore a set of binaural microphones that provides close-talking

http://dx.doi.org/10.21437/Interspeech.2019-2174

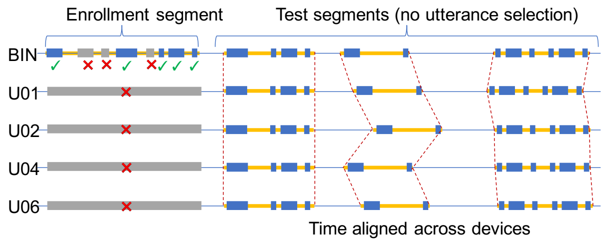

Enrollment segment
Test segments (no utterance selection)
BIN
U01
U02
U04
U06
✓ ✗ ✗ ✓ ✗ ✓ ✓ ✓
Time aligned across devices

Figure 1: *Depiction of the procedure used to extract enrollment and test segments from the CHiME-5 recordings. Audio segments are marked in yellow, and selected speaker utterances in blue. Gray areas represent discarded regions/annotations.*

recordings. All audio was distributed at 16 KHz sampling rate.

For each speaker, a reference transcription was manually produced by listening to the binaural recording. Each utterance was time marked with start and end times and time-aligned for all devices using the binaural recording as reference (see Section 4.1 in [9] for details).

2.2. Derived speaker recognition corpus

The main goals of the design were: i) to create a speaker recognition benchmark in which the performance gap between distant/far-field microphone and a close-talking microphone could be measured; ii) to facilitate research to address the challenges introduced by overlapping multi-speaker conversational speech in noisy and reverberant environments.

To accomplish these goals, we used the 18 parties assigned to the Train and Dev partitions of the CHiME-5 challenge [9]. The remaining 2 parties in the Eval partition were set apart for future use. The two channels of the binaural microphone were summed into a single channel with equal weights. We refer to this device as BIN throughout this paper. We dismissed the arrays U03 and U05 since they had some technical problems in portions of the recordings, and also dropped one of the 40 available speakers (P54) due to an unreliable close-talking microphone. Table 1 summarizes the properties of the derived corpus.

Figure 1 shows the process used to extract enrollment and test segments from the original CHiME-5 recordings. We define "segment" (yellow regions) as the contiguous audio chunk that contains all the selected utterances (blue regions) from the person of interest (POI). Each segment is stored as a separate audio file when we process the CHiME-5 audio to produce our derived corpus.

For enrollment, we only use the audio from the close-talking recordings of their BIN device. For each speaker, we selected the audio from the beginning of the first party they attended. The target amount of speech was set between 40 and 60 seconds. We were highly selective in the utterances picked for enrollment. Based on the time marks of the transcribed utterances, we listened to the audio and only picked utterances with low noise and no speaker overlap. For some of the speakers it took up to one hour of the recording to collect enough speech. This was mostly due to the presence of background noise and simultaneous conversations between the participants. Although this process might seem wasteful, we opted for this approach to minimize the potential performance degradation due to noisy enrollment speech. Figure 2a shows the total duration of the enrollment utterances for each of the 39 speakers in the corpus. The average POI speech-to-segment duration ratio was 5%, which highlights the high selectivity used in the process.

Table 1: *Overview of derived speaker recognition corpus.*

	Enroll	Test
Parties	11	17
Speakers	39	39
Devices	BIN	BIN, U01, U02, U04, U06
Segments	39	5 × 4439

The remaining audio (not used for enrollment) was used for the test partition. For each test segment, five time-aligned versions were extracted using the devices shown in Table 1 (for the microphone arrays each segment comprises 4 audio files). The target amount of speech was set between 10 to 30 seconds. Unlike in the enrollment case, no quality assessment was used to select the utterances to build the segments. That is, given a target amount of desired speech, the segments were built by sequentially including all the transcribed utterances until the duration requirement was met. Figure 2b shows a histogram of the total amount of speech in the generated test segments. A total of 4439 segments were generated from each device. A histogram of the POI to segment duration ratio is shown in Figure 2c. The mean and median values of this ratio are 40 and 36, respectively.

As part of our open source baseline system, we are releasing a script that processes the original CHiME-5 distribution and generates the audio segments and utterance annotations[1].

3. Tasks definitions

We define four tasks based on two enrollment and two test conditions. These conditions mimic those in the SITW corpus [1]. All tasks use the same segments but differ in the way they use the provided time marks for the POI. We note that the task definitions are independent of the approaches used to solve it (e.g., single-microphone vs multi-microphone).

3.1. Enrollment conditions

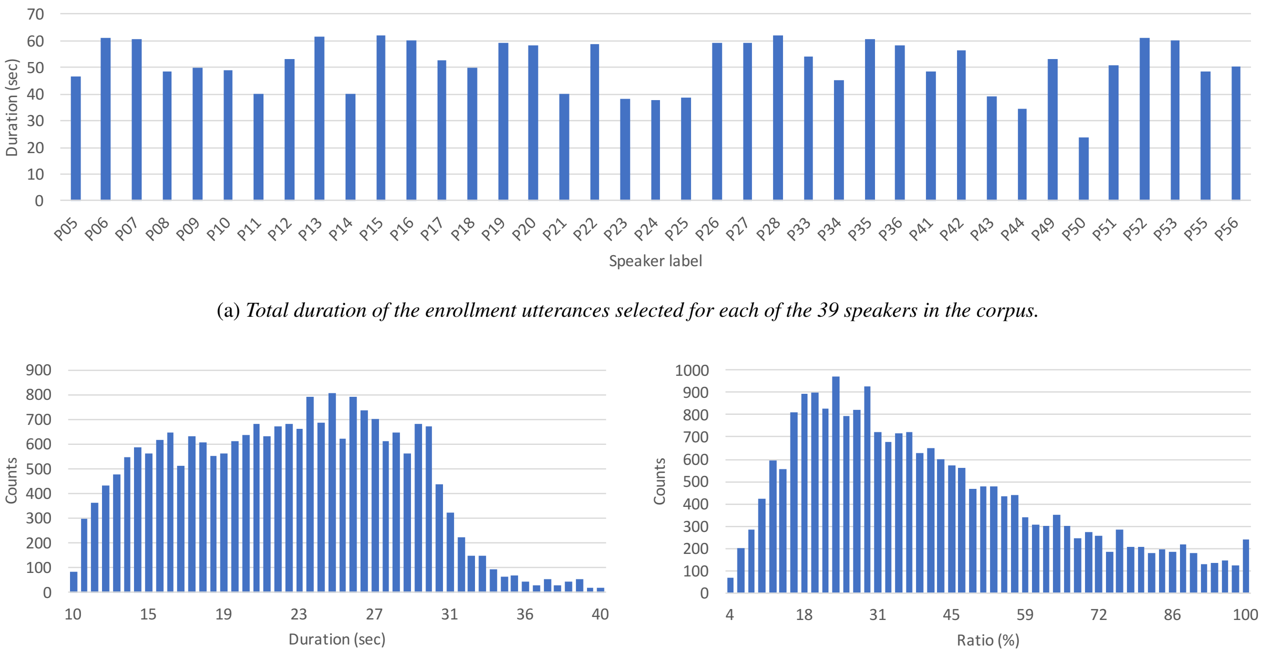
In this benchmark, we only enroll one model per speaker. Therefore, both enrollment conditions result in 39 models. The two conditions are:

1. **Core**: The speaker is enrolled using only the annotated utterances within the audio segment. Figure 2a shows the total duration of the utterances for each speaker.
2. **Assist**: In addition to the annotated utterances, using the entire segment is allowed. Since the enrollment segments are multi-speaker, diarization can find other utterances from the POI.

3.2. Test conditions

1. **Core**: Only the annotated utterances within the test segment are used to compare against a speaker model. Figure 2b shows a histogram of the duration of the annotated utterances.
2. **Multi**: The annotations are ignored and the system needs to handle the presence of multiple speakers. Figure 2c shows a histogram of the ratios of POI speech to test segment duration.

[1] We are currently cleaning up the recipe and plan to make it available in the Kaldi repository during the next few weeks.



(a) *Total duration of the enrollment utterances selected for each of the 39 speakers in the corpus.*

(b) *Histogram of the total amount of POI speech in the test segments.*

(c) *Ratio of POI speech to test segment duration.*

Figure 2: *Statistics of the derived speaker recognition corpus. See Section 2.2 for details.*

3.3. Evaluation protocol

The protocol is the same for all four tasks, and it is designed as a speaker detection problem in which we want to know whether a given POI is present in a test segment or not. Each comparison between an enrolled POI and a test segment is called a trial. If the POI is present in the test segment, we refer to it as a target trial. If it is not present, we call it a non-target trial.

To construct non-target trials, we score each model against all the test segments from any other speaker that does not attend the same parties (to guarantee that potential transcription difficulties do not result in mislabelled non-target trials). This results in a total of 778,025 (5×155,605) non-target trials.

To facilitate performance analysis across devices, we always include all 778,025 non-target trials (using all devices) and break down performance by device of the test segments used for the target trials (4439 per device). In this way, each bar in Figure 3 was computed based on 4439 target trials and 778,025 non-target trials. We also report results averaging the performance across devices.

When processing trials, systems can decide to process each trial independently or using the other enrolled speakers. Moreover, the test segments obtained from the microphone arrays offer the possibility of using multi-microphone approaches (4 mics per array) to process those test segments. When reporting results on these benchmarks, the system descriptions should clearly disclose which one of these available alternatives were used. All the baseline results reported in this paper process each trial independently and use a single-microphone approach.

3.4. Metrics

Systems should report results in terms of equal error-rate (EER) and minimum normalized detection cost (minC) [10] defined by $P_{\text{Target}} = 0.01$, and $C_{FA} = C_{MISS} = 1$.

4. Baseline system

The baseline system uses a speaker recognition component and an agglomerative hierarchical clustering (AHC) diarization component. Both are based on the state-of-the-art DNN embedding (x-vector) described in [11]. Diarization is used for the Assist-enroll and the Multi-test conditions as described in [11].

4.1. Description

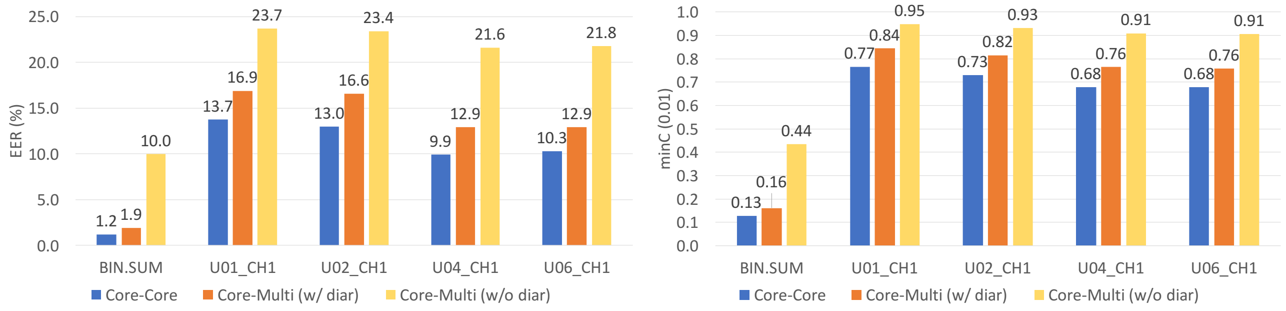
The first layers of the x-vector DNN operate on speech frames, with a small temporal context centered around the current frame t. A pooling layer, aggregates over the input segment, and computes its mean and standard deviation. These segment-level statistics are concatenated together and passed through the remaining layers of the network. The output layer computes posterior probabilities for the training speakers.

The features are 30 dimensional MFCCs with a frame-length of 25 ms, mean-normalized over a sliding window of up to 3 seconds. Audio files are sampled at 16 kHz. The Kaldi energy SAD is used to filter out nonspeech frames.

The system is trained on a large subset of the combined VoxCeleb-1 [3] and VoxCeleb-2 [4] corpora. We removed the overlapping speakers with the SITW corpus, which leaves over 150,000 recordings from 7,185 speakers. We apply data augmentation by adding noises, music, babble, and reverberation [12] . A training example consists of a 2–4 second speech segment along with the corresponding speaker label.

Once the network is trained, x-vectors are extracted from the first affine component after the pooling layer (512 dimensions). The x-vectors are used as features for two different PLDA classifiers: one for the diarization system, and one for the speaker recognition system (both described in [11]).

The PLDA classifier for speaker recognition was trained on the full-length recordings of VoxCeleb. We apply augmenta-



(a) *EER metric*

(b) *minC metric*

Figure 3: *Performance of Core-Core and Core-Multi (with and without diarization of the test segments) tasks broken down by device of the target trials segments. Note that enroll is always based on the BIN device, and the non-target trials are pooled across all devices for each entry in the plot.*

Table 2: *Performance of our baseline system on SITW and (for context) the best submission to the 2016 official evaluation.*

SITW Task	**Best 2016 [2]**		**Baseline [11]**	
Enroll-Test	EER(%)	minC	EER(%)	minC
Core-Core	5.9	0.50	1.7	0.20
Assist-Core	4.5	0.40	1.6	0.20
Core-Multi	7.3	0.57	2.0	0.22
Assist-Multi	5.7	0.46	2.0	0.22

Table 3: *Performance averaged across devices for the tasks defined in the derived CHiME-5 corpus.*

Task	**w/o test diar**		**w/ test diar**	
Enroll-Test	EER(%)	minC	EER(%)	minC
Core-Core	9.6	0.60	–	–
Assist-Core	9.3	0.62	–	–
Core-Multi	20.1	0.83	12.2	0.67
Assist-Multi	18.3	0.80	11.8	0.68

tion to double the amount of training data from about 150,000 to 300,000. Finally, the diarization classifier was trained on 256,000 three-second segments extracted randomly from the full-length augmented recordings.

4.2. Results

Table 2 compares the performance of the baseline system with respect to the best official submission (fusion of multiple systems) to the SITW 2016 evaluation. We show this to provide a context for the strength of the baseline system, which highlights the progress the community has made thanks to the availability of more training data (VoxCeleb corpus) and the advances in DNN architectures that efficiently use it. Additionally, comparing the error rates with Table 3, we can see that the CHiME-5 benchmark is extremely challenging (3x and 6x increase in minC and EER, respectively).

Figure 3 shows the performance of the Core-Core, and Core-Multi (with and without diarization of the test segments) tasks for both metrics. The results are broken down by the device of the target trial segments. Recall that enrollment is always based on the BIN device and that the non-target trials are pooled across all devices to increase the number of trials. For both metrics and tasks, there is a significant degradation when comparing the close-talking BIN microphone performance with the far-field devices. For example, the Core-Core EER increases approximately 10x and the minC around 6x. This shows that there is a great opportunity for noise- and reverberation-robust techniques to have an impact.

Looking at the Core-Multi results, when diarization is not applied to the test segments (yellow bars), we can see that directly detecting a POI in a multi-speaker segment (ignoring the presence of other speakers) produces a significant degradation with respect to the “oracle” diarization of the Core-Core condition. The gap between these two task shows the potential gains available for diarization. Our diarization strategy seems to recover a significant portion of this gap (orange bars); however, the recovered amount is larger for the close-talking BIN microphone than for the far-field devices. This highlights that the artifacts in the distant speech recordings also affect the diarization system.

Finally, Table 3 shows the performance averaged across devices for the four tasks. Ideally, increasing the amount of enrollment data for the Assist-enroll condition has the potential to improve performance. However, we only observed small gains with our baseline system. This indicates that leveraging the additional POI utterances is challenging. Nonetheless, this opens the door for more sophisticated approaches that use speech separation and enhancement to produce benefits.

5. Conclusions

The derived CHiME-5 speaker recognition benchmark is designed to foster robustness against the artifacts introduced by far-field multi-speaker recordings of naturally-occurring spoken interactions. We have presented the process used to build the benchmark and validated its design using a strong state-of-the-art baseline. The observed performance shows that the proposed tasks are extremely challenging. Moreover, the performance gaps between close-talking vs far-field and single-speaker vs multi-speaker recordings provide a clear indication of the potential benefits of robust techniques against noise, reverberation, speech overlap, and multi-speaker recordings.

6. References

[1] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The speakers in the wild (SITW) speaker recognition database," in *Interspeech*, 2016.

[2] M. McLaren, L. Ferrer, D. Castan, and A. Lawson, "The 2016 speakers in the wild speaker recognition evaluation," in *Interspeech*, 2016.

[3] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," in *Interspeech*, 2017.

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

[5] H. Zeinali, H. Sameti, and T. Stafylakis, "Deepmine speech processing database: Text-dependent and independent speaker verification and speech recognition in persian and english," in *Odyssey*, 2018.

[6] A. Janalizadeh Choobbasti, M. Erfan Gholamian, A. Vaheb, and S. Safavi, "JSpeech: A multi-lingual conversational speech corpus," in *SLT*, 2018.

[7] C. Richey, M. A. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. Kumar Nandwana, A. Stauffer, J. van Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices obscured in complex environmental settings (VOiCES) corpus," in *Interspeech*, 2018.

[8] J. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Interspeech*, 2018.

[9] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Interspeech*, 2018.

[10] "The NIST year 2010 speaker recognition evaluation plan," 2010.

[11] D. Snyder, D. Garcia-Romero, G. Sell, A. McCree, D. Povey, and S. Khudanpur, "Speaker recognition for multi-speaker conversations using x-vectors," in *ICASSP*, 2019.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *ICASSP*, 2018.