CNSRC 2022 Evaluation Plan

DONG WANG, Center for Speech and Language Technologies, Tsinghua University, China
QINGYANG HONG, School of Informatics, Xiamen University, China
LANTIAN LI, Center for Speech and Language Technologies, Tsinghua University, China
HUI BU, Beijing Shell Shell Technology Co. Ltd, China

1 INTRODUCTION

The CN-Celeb Speaker Recognition Challenge 2022 (**CNSRC 2022** [1]) aims to evaluate how well the current speaker recognition methods work in real world scenarios, usually with in-the-wild complexity and real-time processing speed. The challenge is based on CN-Celeb, a large-scale free database with the most real-world complexity so far. This document describes the task, performance metric, data, and evaluation protocol of CNSRC 2022.

Compared to other challenges/evaluations such as NIST SRE and VoxSRC, CNSRC 2022 holds the following features:

- It focuses on complex conditions and scenarios, in particular with multi-genre and cross-genre complexity.
- It settles tasks on both speaker verification and large-scale speaker retrieval.

Participation in CNSRC 2022 is open to all people who are interested in speaker recognition technology and are able to comply with the evaluation rules set forth in this plan. There is no cost to participate in CNSRC 2022, however the participants are required to attend the post-challenge workshop held during **Odyssey 2022** [2].

2 TASK DESCRIPTION

CNSRC 2022 defines two tasks: speaker verification (SV) and speaker retrieval (SR). The SV task involves two tracks: *fixed* track and *open* track, and the SR task involves only *open* track. Table 1 summarizes the two tasks.

Table 1. Task definition of CNSRC 2022.

Task	Track	Description	Data	Metric
SV	fixed	Speaker Verification	CN-Celeb	minDCF (EER)
SV	open	Speaker Verification	CN-Celeb + any data	minDCF (EER)
SR	open	Speaker Retrieval	CN-Celeb + any data	mAP

2.1 Speaker Verification

Given a test utterance and a *claimed* speaker, the purpose of the speaker verification task is to automatically determine whether the test utterance is spoken by the claimed speaker. The test utterance along with the enrollment data from the claimed speaker constitutes a *trial*. The system is required to process each trial independently and to output a log-likelihood ratio (LLR), using natural logarithm, for that trial. The LLR for a given trial including a test utterance s is defined as follows:

[1] http://cnceleb.org
[2] http://www.odyssey2022.org

Publication date: January 2022.

$$LLR(s) = \log(\frac{P(s|H_0)}{P(s|H_1)}), \tag{1}$$

where $P(\cdot)$ denotes the probability density function (pdf), and H_0 and H_1 represent the null (i.e., s is spoken by the claimed speaker) and alternative (i.e., s is not spoken by the claimed speaker) hypotheses, respectively. The scoring system will use the LLR values of all the trials to compute the metrics to measure the overall performance.

According to the data used in system development, two tracks are defined for the SV task: *fixed* track and *open* track, shown as follows:

- **Fixed Track**, where only the CN-Celeb.T (ref. Section 3) is allowed for training/tuning the system, including both the front-end and back-end models. No pre-trained front-end and back-end models are permitted. Light-weighted tools such as VAD, speech enhancement, data augmentation could be trained with external data, however the usage must be specified in the system description. This track is designed in order to compare different techniques under the same data resource.
- **Open Track**, where any data sources can be used for developing all the models, except the evaluation set CN-Celeb.E (ref. Section 3). This track is designed to examine the performance frontier of the present technologies with unlimited amount of data.

The desired format of the output file from the SV system is as follows.

Table 2. Format of the SV system output.

<enroll-utt1 ID> <test-utt1 ID> <LLR>
<enroll-utt1 ID> <test-utt2 ID> <LLR>
...
<enroll-uttM ID> <test-uttN ID> <LLR>

For example:

id00800-enroll id00800-singing-01-001 0.9832
id00800-enroll id00802-speech-04-024 0.1031
...
id00999-enroll id00999-singing-02-006 0.7265

2.2 Speaker Retrieval

The purpose of Speaker Retrieval (SR) task is to find out the utterances spoken by a *target* speaker from a large data pool, given an enrollment data of the target speaker. Each target speaker forms a *retrieval request*. The SR system should process each request independently. For each retrieval request, the SR system is required to output the IDs of the hypothesized utterances, and the scoring tool will compute the performance metric, according to the rank of the correct detection.

More specifically, there are S target speakers selected from CN-Celeb.E, and each target individual has 1 enrolled utterance and M test utterances. The non-target set contains a large amount of utterances, coming from multiple sources. The target and non-target utterances are put together, and the participants are required to design their retrieval system to find top-N candidates for each target speaker, and list them in descending order according to the LLR scores. Participants can use any data sources to train their system, **except** CN-Celeb.E.

The desired format of the output file of the SR system is as follows, where the detected utterances for each target speaker are sorted according to the LLR score.

Table 3. Format of the SR system output.

<Speaker-1 ID> <Utterance-1 ID> <Utterance-2 ID> ... <Utterance-10 ID>
<Speaker-2 ID> <Utterance-1 ID> <Utterance-2 ID> ... <Utterance-10 ID>
.....
<Speaker-S ID> <Utterance-1 ID> <Utterance-2 ID> ... <Utterance-10 ID>

3 DATA

3.1 CN-Celeb

CN-Celeb is the main database used in the CNSRC 2022. The database is free and can be downloaded from OpenSLR (http://openslr.org). This database contains two datasets: CN-Celeb1 and CN-Celeb2. The statistics of the two datasets are shown below, and more details can be found in [1, 2]. A key feature of CN-Celeb is that it involves multiple genres (11 genres in total), which makes the intra-speaker variation quite complex.

Table 4. Comparison between *CN-Celeb1* and *CN-Celeb2*.

	CN-Celeb1	CN-Celeb2
Language	Chinese	Chinese
Genre	11	11
# of Sources	1	5
# of Spks	997	1,996
# of Utters	126,532	524,787
# of Hours	271	1,084

A subset of 200 speakers from CN-Celeb1/eval was constituted as the evaluation set, denoted by CN-Celeb.E; all the rest data was used as the training set, denoted by CN-Celeb.T. Note that the definition of CN-Celeb.T and CN-Celeb.E is exactly the same as the training set and evaluation set in Kaldi/cnceleb recipe.

Table 5. Data profile of CN-Celeb.T.

CN-Celeb1/dev	# of Speakers	797
	# of Utters	107,953
CN-Celeb2	# of Speakers	1,996
	# of Utters	524,787
Overall	# of Speakers	2,793
	# of Utters	632,740

Table 6. Data profile of CN-Celeb.E.

Enroll Data	# of Speakers	196
	# of Utters	196
	Avg. Duration	28s
Test Data	# of Speakers	200
	# of Utters	17,777
	Avg. Duration	8s
Trials	# of Target	17,755
	# of Non-target	3,466,537

3.2 Data profile for speaker verification

For the SV task, CN-Celeb.E is used as the evaluation set in both the fixed track and open track.

In the *fixed* track, only CN-Celeb.T is allowed to be used to perform system development, including all the major components: front-end embedding model, back-end scoring model, normalization/calibration. Pre-trained models and off-the-shelf tools can be used for other components, e.g., VAD and speech enhancement. Data augmentation is allowed, but only limited to signal processing approaches (noise mixing, speed perturbation, reverberation simulation) with public noise datasets. Data augmentation based on models trained with extra data resources is not allowed, e.g., TTS-based augmentation. All the extra tools and augmentation methods used for system development should be clearly stated in the system description.

In the *open* track, any data sources and tools can be used to develop the system, though the used tools and data should be clearly stated in the system description.

3.3 Data profile for speaker retrieval

For the SR task, two datasets will be released: SR.dev and SR.eval. Each dataset contains two parts: (1) Target speakers and associated enrollment data; (2) Utterance pool that involves utterances of the target speakers as well as a large amount of non-target utterances. SR.dev will be provided to the participants for system development, while SR.eval will be released for system evaluation.

The target speakers in both SR.dev and SR.eval are selected from CN-Celeb.E, and the speakers in the two sets are not overlapped. Specifically, SR.dev consists of 5 target speakers, each with 10 test utterances; and SR.eval consists of 25 target speakers, each with 10 test utterances. Note that the enrollment and test utterances could be different from the data split of CN-Celeb.E in Kaldi/cnceleb recipe.

The non-target utterances are selected from CN-Celeb.T. The duration of these utterances varies from a few seconds to tens of seconds, and the acoustic conditions are varied and complex. The number of non-target utterances is $20,000$ in SR.dev and $500,000$ in SR.eval.

4 PERFORMANCE MEASUREMENT

The metrics used for performance evaluation are described in this section. The toolkit used to compute these metrics has been integrated in the code of the baseline systems.

4.1 Speaker Verification

The primary metric for SV performance evaluation is *minimum Detection Cost Function (minDCF)*.

Firstly define the detection cost function as follows:

$$C_{Det}(\theta) = C_{Miss} \times P_{Target} \times P_{Miss}(\theta) + C_{FalseAlarm} \times (1 - P_{Target}) \times P_{FalseAlarm}(\theta), \quad (2)$$

where $P_{Miss}(\theta)$ is the missing rate and $P_{FalseAlarm}(\theta)$ is the false alarm rate with the decision threshold set to θ. C_{Miss} and $C_{FalseAlarm}$ are the cost of a missed detection and a spurious detection, respectively; P_{Target} is a prior probability of the specified target speaker. Then *minDCF* is obtained by minimizing $C_{Det}(\theta)$ with respect to θ and setting $C_{Miss} = C_{FalseAlarm} = 1$ and $P_{Target} = 0.01$:

$$minDCF = \arg\min_{\theta} \{0.01 \times P_{Miss}(\theta) + 0.99 \times P_{FalseAlarm}(\theta)\}. \quad (3)$$

Besides *minDCF*, the SV performance is also evaluated/analyzed in three ways:

- *Equal Error Rate (EER).* EER is defined as the balanced value of P_{Miss} and $P_{FalseAlarm}$, formally $P_{Miss}(\theta^*) = P_{FalseAlarm}(\theta^*)$, where θ^* is the decision threshold that achieves the balance. EER is used as the *auxiliary* metric and should be reported in the system description.
- *Decision Error Tradeoff (DET) curve.* DET curve is a curve within a two-dimensional space where the two axes represent P_{Miss} and $P_{FalseAlarm}$ respectively. The DET curve reflects the trade-off between missing and false alarm, and presents the performance of the system at various operation points determined by θ.

4.2 Speaker Retrieval

The performance of the SR system will be measured in terms of *Mean Average Precision (mAP)*.

For a single speaker i, suppose there are M test utterances overall, and the system output maximum top-N candidates for each retrieval request. For the top-k case, the *Precision* is defined as:

$$Precision(i,k) = \frac{\sum_{j=1}^{k} \delta(utt_j \; is \; from \; speaker \; i)}{k}. \quad (4)$$

The AP of top-N is defined as the averaged precision over the top-k ($k = 1, 2, .., N$) cases:

$$AP(i) = \frac{1}{N} \sum_{k=1}^{N} Precision(i,k). \quad (5)$$

Then mAP is computed as the averaged AP over all the target speakers:

$$mAP = \frac{1}{S} \sum_{i=1}^{S} AP(i), \quad (6)$$

where S is the number of target speakers. For the evaluation set of CNSRC 2022, the parameters are as follows: the number of target speakers $S = 5$ in SV.dev and $S = 25$ in SV.eval, the number of test utterances per target speaker $M = 10$, and the SR system output maximum candidates $N = 10$.

5 EVALUATION PROTOCOL

5.1 Registration

Participants must sign up for an evaluation account where they can perform various activities such as registering for the evaluation, signing the data license agreement, as well as uploading the submission and system description. To sign up for an evaluation account, go to http://cnceleb.org/competition.

Once the account has been created, the registration can be performed online. The registration is free to all individuals and institutes. The regular case is that the registration takes effect immediately,

but the organizers may check the registration information and ask the participants to provide additional information to validate the registration.

5.2 Baselines

The organizers prepared multiple baseline systems to demonstrate the process of training/evaluation required by the challenge. All the baseline systems are open-sourced.

5.2.1 Speaker Verification. For the speaker verification task, since the data profile and evaluation protocol are the same as the Kaldi-based recipe [3], this recipe can be used directly as the baseline, in particular for the fixed track. This recipe can be easily adapted to develop an open-track system by involving more training data.

Besides Kaldi, the organizers additionally built two PyTorch baselines:

- CNSRC 2022 recipe in ASV subtools [3]. ASV subtools is an open source toolkit for speaker recognition and language recognition, published by Xiamen University. Participants can clone the code from github [4] and use the cnsrc/sv recipe to reproduce the fixed-track SV system. Open-track systems can be easily trained by using the same recipe but with more data.
- CN-Celeb recipe in Sunine toolkit [5]. Sunine is an open source toolkit for speaker recognition, with terse implementation of several SOTA techniques, published by Tsinghua University. The cnceleb/v2 recipe can be used to reproduce the fixed-track SV system. With more data, an open-track system can be built using the same recipe.

5.2.2 Speaker Retrieval. For speaker retrieval, the organizers published two baseline systems with the ASV subtools and Sunine toolkit, respectively. The recipe is cnsrc/sr in the ASV subtools, and cnceleb/v3 in the Sunine toolkit.

5.3 Result Submission

Participants should submit their results via the submission system. For the SV task, the submission is the output file of the SV system tested on CN-Celeb.E according to the format shown in Table 2. For the SR task, the submission is the output file of the SR system tested on SR.eval according to the format shown in Table 3.

Once the submission is completed, it will be shown in the *Leaderboard*, and all participants can check their positions. For each task and each track, participants can submit their results no more than 10 times.

5.4 System Description

Each participant is required to submit a system description. The system description must include the following items:

- A complete description of the system components, including front-end (e.g., VAD, diarization, features, embedding model) and back-end (e.g., LDA/PLDA, normalization, calibration, retrieval model) modules along with their configurations (i.e., filter bank configuration, dimension and type of the acoustic feature parameters, as well as the configurations of embedding model, scoring model and retrieval model).

[3] https://github.com/kaldi-asr/kaldi/tree/master/egs/cnceleb
[4] https://github.com/Snowdar/asv-subtools
[5] https://gitlab.com/csltstu/sunine

- A complete description of the data profile used to train the various models (as mentioned above). Participants are encouraged to report how different training schemes impact the performance, and how systems are combined.
- Performance of the submission systems. For the SV task, the participants can compute their minDCF and EER scores on the CN-Celeb.E set according to the SV baselines. For the SR task, the participants should report the results on the SR.dev set computed according to the SR baselines as well as the results on the SR.eval set returned by the submission system. Participants are encouraged to quantify the contribution of their major system components that they believe resulted in significant performance gains.
- A report of the CPU (single thread) and GPU execution times as well as the amount of memory used to process a single trial in the SV task (i.e., the time and memory used for creating a speaker model from enrollment data as well as processing a test segment to compute the LLR) and a single retrieval request in the SR task (i.e., the time and memory used for creating the speaker model from the enrollment data as well as searching the target speaker among the utterance pool).

The system description should follow the template presented in the Odyssey 2022 [6].

5.5 Paper Submission

The participant is encouraged to submit the system description as an Odyssey 2022 paper. The deadline for the paper submission is later than the deadline for regular papers, but the review process is identical. The challenge papers will be treated the same as regular papers and will be presented in the post-evaluation workshop in Odyssey 2022.

5.6 Workshop

The post-evaluation workshop will be held as a special event in Odyssey 2022. The participants are required to attend the workshop. Authors of high-rank systems will deliver reports, and accepted challenge papers will be presented as well.

6 TIME SCHEDULE

Mid Feb	Registration system open
Late Feb	Development set for SR task release
Mid Mar	Evaluation set for SR task release
Mid Mar	Submission system and Leaderboard open
Mid May	Deadline for submission of results
Late May	Deadline for system description and challenge paper
Mid Jun	Challenge paper notification
29th Jun	CNSRC 2022 workshop at Odyssey 2022

REFERENCES

[1] Yue Fan, JW Kang, LT Li, KC Li, HL Chen, ST Cheng, PY Zhang, ZY Zhou, YQ Cai, and Dong Wang. 2020. CN-CELEB: a challenging Chinese speaker recognition dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7604–7608.

[2] Lantian Li, Ruiqi Liu, Jiawen Kang, Yue Fan, Hao Cui, Yunqi Cai, Ravichander Vipperla, Thomas Fang Zheng, and Dong Wang. 2020. CN-Celeb: multi-genre speaker recognition. *arXiv preprint arXiv:2012.12468* (2020).

[3] Fuchuan Tong, Miao Zhao, Jianfeng Zhou, Hao Lu, Zheng Li, Lin Li, and Qingyang Hong. 2021. ASV-Subtools: Open Source Toolkit for Automatic Speaker Verification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6184–6188.

[6] http://www.odyssey2022.org/col.jsp?id=119