CNVSRC 2025 Evaluation Plan

DONG WANG, Center for Speech and Language Technologies, Tsinghua University, China
LANTIAN LI, Beijing University of Posts and Telecommunications, China
ZEHUA LIU, Beijing University of Posts and Telecommunications, China
XIAOLOU LI, Beijing University of Posts and Telecommunications, China
KE LI, Beijing Haitian Ruisheng Science Technology Ltd., China
HUI BU, Beijing AIShell Technology Co. Ltd., China

1 INTRODUCTION

Visual speech recognition (VSR), also known as lip reading, is a technology that infers pronunciation content through lip movements. A related task, known as visual to speech conversion (VTS), set up its purpose as converting lip movement to speech signals directly. These two tasks, VSR and VTS, share the same key issues and technical solutions.

Currently, research on VSR is still in its early stages and cannot accommodate real-life applications. Significant progress has been made in recongizing/converting isolate words/phrases, especially in speaker-dependent scenarios. However, for spontaneous speech with unconstrained vocabulary and unconstrained speakers, the performance with VSR and VTS is still poor. Especially for Chinese, research progress is greatly constrained due to the lack of relevant data resources. In 2023, CSLT@Tsinghua University released the CN-CVS dataset[1], becoming the first large-scale Chinese visual-speech multi-modal database, providing possibilities for further promoting large vocabulary continuous visual speech recognition (LVCVSR) and related tasks, e.g., large vocabulary continuous visual to speech conversion (LVCVTS).

To expand this important research direction, CSLT@Tsinghua University, together with Beijing University of Posts and Telecommunications, Beijing Haitian Ruisheng Science Technology Ltd., and Speech Home, kicked off the first Chinese Continuous Visual Speech Recognition Challenge in 2023 (CNVSRC 2023) [1] and continue to hold CNVSRC 2024 in the next year [2]. The participants contributed a lot of new ideas and new technologies.

CNVSRC 2025[2] keeps the research focus on visual information processing and is a special event at the NCMMSC 2025 conference[3].

Compared to CNVSRC 2023 and CNVSRC 2024, CNVSRC 2025 is characterized by two important changes: (1) introduce a new visual to speech (VTS) track (2) release of additional 1,000 hours of training data, to support large-scale models. The organizers provide baseline codes[4] to refer to participants. The results of CNVSRC 2025 will be announced and awarded at NCMMSC 2025.

2 DATA

· *CN-CVS*: CN-CVS contains visual-speech data from over 2,557 speakers with more than 300 hours of data, covering news broadcasts and public speaking scenarios, and is currently the largest open-source Chinese visual-speech dataset. The organizers have provided text annotations of this database for this challenge. For more information about CN-CVS, please visit its official website (*http://www.cnceleb.org/*).

[1]http://cnceleb.org/
[2]http://cnceleb.org/competition
[3]http://www.ncmmsc.org.cn/
[4]https://github.com/liu12366262626/CNVSRC2025

Publication date: July 2025.

· *CNVSRC-Single*: CNVSRC single-speaker data. It includes audio and video data from a single speaker with over 90 hours of data, obtained from internet media. The total data consists of 90 hours (25,947 samples) as the development set, and approximately 1 hour (300 samples) of data as the evaluation set.
· *CNVSRC-Multi*: CNVSRC multi-speaker data. It includes audio and video data from 43 speakers, with nearly 1 hour of data per person. Two-thirds of each person's data make up the development set, while the remaining data make up the evaluation set. The data from 23 speakers were recorded in a recording studio with fixed camera positions and reading style, and each recording is relatively short. The data from the other 20 speakers were obtained from internet speech videos, with longer recording duration and more complex environments and content.
· *CN-CVS2-P1*: An extra data source for the fixed tracks. For the fixed tracks, the organizer provides an extra data source, a preview part of the CN-CVS2 dataset, for system development. It encompasses over 160,000 utterances with a total duration of about 200 hours sourced from internet media.
· *CN-CVS3*: This is an additional data source specifically provided for this year's competition. It's a significant expansion, offering over 900,000 utterances with a total duration of approximately 1000 hours. This extensive visual-speech dataset, primarily sourced from internet media, is designed to serve as a comprehensive resource for system development and training.

In the baseline training process, we simultaneously utilized the aforementioned datasets.

Table 1. Data profile of CNVSRC-Single, CNVSRC-Multi, CN-CVS2-P1 and CN-CVS3.

	CNVSRC-Single		CNVSRC-Multi		CN-CVS2-P1	CN-CVS3
DataSet	Dev	Eval	Dev	Eval	Train	Train
# Videos	25,947	300	20,450	10,269	160,116	900,001
# Hours	94.00	0.87	29.24	14.49	196.77	990.45

For the training and development sets, the organizers provide audio, video, and corresponding transcribed text. For the evaluation set, only video data will be provided. Participants are prohibited from using the evaluation set in any way, including but not limited to using the evaluation set to help train or fine-tune their models.

Note: The reading data in CNVSRC-Multi comes from the dataset[5]. This dataset was donated to CSLT@Tsinghua University by Beijing Haitian Ruisheng Science Technology Ltd.[6] to promote scientific development.

3 TASK AND TRACK

CNVSRC 2025 consists of two tasks: Multi-speaker VSR (T1) and Single-speaker VTS (T2). The former (T1) focuses more on the accuracy of content recognition for multiple speakers, while the latter (T2) emphasizes the restoration of content and speaker audio features under the condition of a single speaker. Each task is divided into 'fixed track' and 'open track', with the fixed track only allowing the use of data and other resources agreed upon by the organizing committee, while the open track can use any resources except the evaluation set.

Specifically, resources that cannot be used in the fixed track include: non-public pre-training models used as feature extractors, pre-training language models with more than 1B parameters, or that

[5]https://dataoceanai.com/dataset/c61-6344.htm
[6]http://dataoceanai.com

Table 2. Task description of CNVSRC 2025.

	Fixed Track	Open Track
T1: Multi-speaker VSR	CN-CVS, CNVSRC, CN-CVS2-P1, CN-CVS3	No constraint
T2: Single-speaker VTS	CN-CVS, CNVSRC, CN-CVS2-P1, CN-CVS3	No constraint

are non-public. Tools and resources that can be used include: publicly available pre-processing tools such as face detection, extraction, lip area extraction, contour extraction, etc.; publicly available external models and tools, datasets for data augmentation; publicly available word lists, pronunciation dictionaries, n-gram language models, and neural language models with less than 1B parameters.

4 EVALUATION PROTOCOL

4.1 Registration

Participants must register for a CNVSRC account where they can perform various activities such as signing the data user agreement as well as uploading the submission and system description. To register for a CNVSRC account, please go to *http://cnceleb.org/competition*.

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.

Once the account has been created, participants can apply the data, by signing the data agreement and uploading it to the system. The organizers will review the application, and if it is successful, participants will be notified the link of the data.

4.2 Baselines

The organizers construct baseline systems for the Multi-speaker VSR task and the Single-speaker VTS task, using the data resource permitted on the *fixed tracks*. The baseline leverages advanced methods for VSR and VTS and offer reasonable performance, as shown below:

Table 3. Performance of baseline systems on the fixed tracks.

Task	Multi-speaker VSR	Single-speaker VTS
CER on Dev Set	31.91%	33.15%
CER on Eval Set	31.55%	31.41%

Participants can download the source code of the baseline systems from *https://github.com/liu12366262626/CNVSRC2025*.

4.3 Result Submission

We use Character Error Rate (CER) as the evaluation metric for both the Multi-speaker VSR and Single-speaker VTS tasks. To submit the results, the participant should sign in to the CNVSRC official website, finding the submission page, selecting the task and track, and then submitting the result file.

For the Multi-speaker VSR task, the file format is simple: each line contains the ID of a video and the hypothesized words.

For the Single-speaker VTS task, participants are required to submit the corresponding generated audio in .wav format for each video ID.

The system will return the CER immediately. The participant can submit no more than 5 times for each task in each track.

All valid submissions are required to be accompanied by a system description, submitted via the submission system. All the system descriptions will be published on the web page of the CNVSRC 2025 workshop.

In the system description, participants are allowed to hide their names and affiliations.

4.4 Workshop

The post-evaluation workshop will be held as a special event in NCMMSC 2025. Authors of high-rank systems will deliver reports. All the participants are welcome to attend the workshop.

5 TIME SCHEDULE

2025/07/04	Registration kick-off
2025/07/04	Data release
2025/07/04	Baseline system release
2025/08/01	Submission system open
2025/10/01	Deadline for result submission
2025/10/16-19	Workshop at NCMMSC 2025

REFERENCES

[1] Chen Chen, Zehua Liu, Xiaolou Li, Lantian Li, and Dong Wang. 2024. Cnvsrc 2023: The first chinese continuous visual speech recognition challenge. In *Interspeech 2024*.

[2] Zehua Liu, Xiaolou Li, Chen Chen, Lantian Li, and Dong Wang. 2025. Cnvsrc 2024: The second chinese continuous visual speech recognition challenge. In *Interspeech 2025*.