

# CNVSRC 2023 Evaluation Plan

DONG WANG, Center for Speech and Language Technologies, Tsinghua University, China

CHEN CHEN, Center for Speech and Language Technologies, Tsinghua University, China

LANTIAN 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, also known as lip reading, is a technology that infers pronunciation content through lip movements. It has important applications in public safety, assisting the elderly and the disabled, and fake video detection. Currently, research on lip reading is still in its early stages and cannot accommodate real-life applications. Significant progress has been made in phrase recognition, but it still faces great challenges in large vocabulary continuous recognition. 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<sup>1</sup>, becoming the first large-scale Chinese visual-speech multi-modal database, providing possibilities for further promoting large vocabulary continuous visual speech recognition (LVCVSR).

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, will hold the **Chinese Continuous Visual Speech Recognition Challenge (CNVSRC) at the NCMMSC 2023 conference**<sup>2</sup>. The organizers will use the CN-CVS dataset as the basic training data, and will test the performance of LVCVSR systems in two scenarios: reading in a recording studio and speech on the Internet. The organizers will provide baseline codes for participants to refer to. The results of CNVSRC will be announced and awarded at NCMMSC 2023.

## 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/>). This dataset will serve as the training set for the fixed tracks of the challenge.

· *CNVSRC-Single*: CNVSRC single-speaker data. It includes audio and video data from a single speaker with over 100 hours of data, obtained from internet media. Nine-tenths of the data will make up the development set, while the remaining one-tenth will serve 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.

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

<sup>1</sup><http://cnceleb.org/>

<sup>2</sup><http://www.ncmmsc.org.cn/>

Table 1. Data profile of CNVSRC-Single and CNVSRC-Multi.

	CNVSRC-Single		CNVSRC-Multi	
DataSet	Dev	Eval	Dev	Eval
# Videos	25,947	2,881	20,450	10,269
# Hours	94.00	8.41	29.24	14.49

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<sup>3</sup>. This dataset was donated to CSLT@Tsinghua University by Beijing Haitian Ruisheng Science Technology Ltd.<sup>4</sup> to promote scientific development.

### 3 TASK AND TRACK

CNVSRC 2023 consists of two tasks: Single-speaker VSR (T1) and Multi-speaker VSR (T2). The former T1 focuses on the performance of large-scale tuning for a specific speaker, while the latter T2 focuses on the basic performance of the system for non-specific speakers. 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.

Table 2. Task description of CNVSRC 2023.

	Fixed Track	Open Track
T1: Single-speaker VSR	CN-CVS, CNVSRC-Single.Dev	No constraint
T2: Multi-speaker VSR	CN-CVS, CNVSRC-Multi.Dev	No constraint

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 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, 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.

<sup>3</sup><https://dataoceanai.com/dataset/c61-6344.htm>

<sup>4</sup><http://dataoceanai.com>

Once the account has been created, participants can apply the data, by signing the data agreement and upload 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 Single-speaker VSR task and the Multi-speaker VSR task, using the data resource permitted on the fixed track. The baselines use the Conformer structure as the building blocks and offer reasonable performance, shown below:

Table 3. Performance of baseline systems.

Task	Single-speaker VSR	Multi-speaker VSR
CER on Dev Set	48.57%	58.77%
CER on Eval Set	48.60%	58.37%

Participants can download the source code of the baseline systems from <https://github.com/MKT-Dataoceanai/CNVSRC2023Baseline>.

## 4.3 Result Submission

CNVSRC 2023 uses Character Error Rate (CER) as the main metric to evaluate the performance. 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.

The file contains format is simple: each line contains the ID of a video and the hypothesized words. 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 with a system description, submitted via the submission system. All the system descriptions will be published at the web page of the CNVSRC 2023 workshop.

In the system description, participants are allowed to hide their name and affiliation.

## 4.4 Workshop

The post-evaluation workshop will be held as a special event in NCMMS 2023. Authors of high-rank systems will deliver reports. All the participants are welcome to attend the workshop.

## 5 TIME SCHEDULE

2023/09/20	Registration kick-off
2023/09/20	Training data, development data release
2023/09/20	Baseline system release
2023/10/10	Evaluation set release
2023/11/01	Submission system open
2023/12/01	Deadline for result submission
2023/12/09	Workshop at NCMMS 2023