Interspeech 2024
1-5 September 2024, Kos, Greece

CNVSRC 2023: The First Chinese Continuous Visual Speech Recognition Challenge

Chen Chen[1], *Zehua Liu*[2], *Xiaolou Li*[2], *Lantian Li*[2], *Dong Wang*[1]

[1]Center for Speech and Language Technologies, BNRist, Tsinghua University, China
[2]Beijing University of Posts and Telecommunications, China

{chenc21,wangdong99}@mails.tsinghua.edu.cn, {lixiaolou,liuzehua,lilt}@bupt.edu.cn

Abstract

The first Chinese Continuous Visual Speech Recognition Challenge aimed to probe the performance of Large Vocabulary Continuous Visual Speech Recognition (LVC-VSR) on two tasks: (1) Single-speaker VSR for a particular speaker and (2) Multi-speaker VSR for a set of registered speakers. The challenge yielded highly successful results, with the best submission significantly outperforming the baseline, particularly in the single-speaker task. This paper comprehensively reviews the challenge, encompassing the data profile, task specifications, and baseline system construction. It also summarises the representative techniques employed by the submitted systems, highlighting the most effective approaches. Additional information and resources about this challenge can be accessed through the official website at *http://cnceleb.org/competition*.

Index Terms: CNVSRC, Visual speech recognition, Lip reading, Chinese VSR

1. Introduction

Visual Speech Recognition (VSR), commonly called lip reading, is a technology that utilizes lip movements to infer speech content. It has broad applications, including public surveillance, support for elderly and disabled individuals, and fake video detection. Traditionally, VSR has been primarily focused on recognizing isolated words or phrases. For example, Martinez et al. [1] developed a model that extracts visual features using 3D convolution, ResNet-18, and a multi-scale temporal convolutional network (MS-TCN). This was further enhanced by simple average pooling and a softmax layer for inferring word posteriors, resulting in commendable performance on the LRW [2] and LRW-1000 datasets [3], which are the largest publicly available benchmark datasets for unconstrained isolated word lip-reading in English and Mandarin, respectively. Ma et al. [4] adopted a similar architecture but introduced a densely connected temporal convolutional network (DC-TCN) to achieve improved performance.

Recently, research in lip reading has progressed beyond word and phrase recognition to focus on large vocabulary continuous visual speech recognition (LVC-VSR), a more challenging task but with more realistic merit. Significant advancements have been made in English benchmarks, partly attributable to the availability of large-scale English visual-speech datasets such as LRS2 [5], LRS3 [5], AVSpeech [6], and VoxCeleb [7, 8]. While earlier studies addressed this issue using hand-crafted features and sequential models like HMM [9] or RNN [10], a significant breakthrough occurred with the end-to-end approach, which processes raw video frames and generates a word sequence. LipNet [11] is perhaps the first end-to-end model, integrating spatiotemporal convolution layers and bidirectional gated recurrent unit (BGRU), trained using the connectionist temporal classification(CTC) loss [12]. The model underwent testing on GRID, a dataset with limited grammar and vocabulary [13]. A similar architecture was adopted by Jeon et al. [14], with their approach involving the integration of multiple CNN-based streams into the feature extraction process.

While promising, the GRID corpus is limited in its ability to capture real-world complexity due to the constrained grammar and vocabulary of the sentences. A more challenging task involves large vocabulary continuous visual speech recognition based on datasets gathered from online media repositories like YouTube. A seminal work [5]introduced the first comprehensive visual speech datasets LRS2 and LRS3, along with the first transformer-based system trained with either the CTC loss (TM-CTC) or the sequence-to-sequence loss (TM-seq2seq). In a parallel endeavour, Google [15] developed an end-to-end model (CNN/BLSTM backbone and CTC loss) that transcribes videos into phone sequences and utilizes an FST-based decoder to obtain word sequences. The research was expanded upon in a subsequent study [16], which introduced a hybrid CTC/Attention model utilizing a ResNet/Conformer encoder and a Transformer-based language model (LM). This work was further developed by incorporating time-masking data augmentation and an auxiliary reconstruction loss in a subsequent publication [17].

In addition to employing complex structures, a simple yet effective approach is increasing the training data volume. However, a major challenge arises from the dearth of well-labeled data. One promising strategy to address this issue involves utilizing a pre-trained automatic speech recognition (ASR) model to transcribe unlabelled videos. This methodology was extensively explored in [18], wherein a ResNet/Conformer model was trained on both unlabelled datasets like VoxCeleb2 [8] and AVSpeech [6], as well as text-labelled datasets such as LRW [2], LRS2 [5], and LRS3 [5]. The study revealed significant performance enhancements with auto-labelled data, and further improvements were observed with increased incorporation of unlabelled data.

For VSR on Chinese data, the research progress has been significantly impeded by the scarcity of data resources. Despite the presence of LRW-1000 [3] as the sole large-scale dataset in Chinese, it consists solely of isolated words. In 2023, the release of the CN-CVS dataset [19] marked the debut of the first large-scale continuous visual-speech dataset in Chinese, thereby presenting an opportunity to propel research in Chinese

This work was supported by the National Natural Science Foundation of China (NSFC) under Grants No.62301075/62171250. L. Li and D. Wang are the corresponding authors.

10.21437/Interspeech.2024-2509

LVC-VSR. This also motivated the first Chinese Continuous Visual Speech Recognition Challenge (CNVSRC 2023), hosted as a special session on NCMMSC 2023 to estimate the performance boundary of existing LVC-VSR techniques with Chinese data and attracting research interest in this domain. The CN-CVS [19] dataset was used as the primary training data, supplemented by two additional datasets - CNVSRC-Single and CNVSRC-Multi - introduced by the organizers to facilitate system development and evaluation. The organizers also published the model and code of the baseline systems so that the participants could use them as references when developing their systems. This paper summarises the challenge, emphasizing the predominant findings gleaned from the submitted systems.

The structure of the rest of this paper is outlined as follows: Section 2 introduces the tasks and data of the challenge. In Section 3, a comprehensive description of the baseline system is provided, covering the model structure, training strategies, and performance evaluation. Section 4 reports the challenge result and summarizes the representative technologies utilized by the participants. Lastly, the paper is concluded in Section 5.

2. Tasks and Data

2.1. Tasks

The CNVSRC 2023 challenge encompasses two distinct tasks: Single-speaker VSR (T1) and Multi-speaker VSR (T2). Task T1 emphasizes the performance of large-scale tuning for a specific speaker, whereas T2 focuses on the fundamental performance of the system for non-specific but *registered* speakers, i.e., speakers seen in the data for system development. In both tasks, the system is fed with silent facial videos featuring a single individual, and it is required to generate the spoken content in written form.

Each task is further categorized into a 'fixed track' and an 'open track'. The fixed track permits the use of data and additional resources that have been agreed upon by the organizing committee. Conversely, the open track allows participants to employ any resources except the evaluation set.

Character Error Rate (CER) was used as the main metric to evaluate VSR performance, formulated as follows:

$$CER = \frac{S + D + I}{N} \tag{1}$$

where S, D, and I represent the number of substitutions, deletions, and insertions in the output transcription, respectively. N is the number of characters in the ground truth transcription.

2.2. Data profile

CNVSRC 2023 utilized the CN-CVS dataset [19] in addition to two supplementary datasets: CNVSRC-Single and CNVSRC-Multi, which served as the development and evaluation data for the single-speaker VSR task (T1) and multi-speaker VSR task (T2), respectively. All the data was transcribed into text. Table 1 presents the data profile of the three datasets. Note that both CNVSRC-Single and CNVSRC-Multi were split into a development set and an evaluation set. The development data was transparent to the participants (including the video, audio, and text), while the text and audio of the evaluation data were kept secret during the entire challenge process. The participants could use the development data in any way, e.g., freely splitting it into a subset for model fine-tuning and a subset for model validation/selection.

Table 1: *Data profile used in CNVSRC 2023.*

	CN-CVS	CNVSRC-Single		CNVSRC-Multi	
DataSet	Train	Dev	Eval	Dev	Eval
# Spks	2,557	1		43	
# Videos	206,261	25,947	2,881	20,450	10,269
# Hours	308.00	94.00	8.41	29.24	14.49

2.2.1. CN-CVS

The CN-CVS dataset [19] comprises visual-speech data from over 2,557 speakers, totalling more than 300 hours of videos. It encompasses various scenarios, including news broadcasts and public speeches. So far, CN-CVS is the largest open-source Chinese visual-speech dataset. This dataset was used as the primary training data for the CNVSRC 2023 challenge. Note that the original publication of CN-CVS is for the video-to-speech synthesis (VTS) task and thus does not involve text transcription. To support the CNVSRC 2023 challenge, we labelled the video with a semi-automatic pipeline that involves ASR transcribing and human check, as will be presented shortly.

2.2.2. CNVSRC-Single

The CNVSRC-Single dataset, designed for a single-speaker VSR task (T1), is obtained from a broadcaster's online channel. It includes over 800 speech videos of that broadcaster, with a cumulative duration of over 100 hours. The pipeline used in [19] was employed for the collection and processing.

2.2.3. CNVSRC-Multi

The CNVSRC-Multi dataset was designed as the development/evaluation data for the multi-speaker VSR task (T2). It encompasses two scenarios: reading in a recording studio and speeches downloaded from the internet.

In the recording studio scenario, facial videos of speakers were captured from three different camera angles (0°,30°,60°), while their speech audio was recorded using a high-quality microphone. The speakers were prompted a sentence at each time via a computer screen and were asked to read the sentence clearly with a neutral emotion. The video data from the *front camera* was used in CNVSRC 2023, which was transcoded to 25 frames per second (FPS) and scaled to an appropriate size. A total of 23 speakers participated in the recording, each reading 1,000 sentences. In the speeches from the internet scenario, videos of public speeches from 20 speakers were collected from the internet, again following the same collection and processing pipeline as [19]. To ensure the integrity of the collected videos from the internet, a face recognition tool[1] was employed to check whether there is only one face in each video frame and if the face is the target face. A manual check was then conducted to double-check that each extracted video only contained the target face.

2.3. Text annotation

To generate text transcriptions, a Paraformer-based ASR system [20] was employed to transcribe the speech of all the videos. Furthermore, the manual check was conducted to ensure that the CER of the transcriptions remains below 2%.

3. Baseline System

Leveraging the state-of-the-art model used in Auto-AVSR [18], we trained two baseline systems, one for the single-speaker

[1] https://pypi.org/project/face-recognition/

Table 2: *Training Details of the Pretraining (P1 & P2) and Fine-tuning (FT) steps when constructing the baseline systems.*

Experiment	P1	P2	FT (Single-Speaker)	FT (Multi-Speaker)
Initialize	Random	P1 Saved Model	P2 Saved Model	P2 Saved Model
Warmup Epochs	5	5	2	2
Learning Rate	0.0002	0.001	0.0003	0.0002
Training Epochs	75 + Early stop	75	80	80
Saved Model	Top 10 average	Last 10 epochs average	Last 5 epochs average	Last 5 epochs average

VSR task (T1) and the other for the multi-speaker VSR task (T2). Only the datasets provided in this challenge were used. In other words, these systems conform to the specifications of the fixed tracks.

3.1. Model structure

The model structure is duplicated from Auto-AVSR [18]. Specifically, it comprises three components: visual frontend, encoder, and decoder. As in [18], the visual frontend adopts ResNet18 as its backbone, except that the first 2D-CNN layer is replaced with a 3D-CNN layer to capture local spatiotemporal correlation. The encoder adopts a Conformer structure with 12 layers, while the decoder employs a Transformer structure with 6 layers.

Upon receiving the input video data, the visual frontend performs initial local spatiotemporal feature extraction. Subsequently, the Conformer encoder further extracts context-dependent features. A projection layer and a transformer decoder are employed to predict the output class labels. The entire model was trained with the joint CTC/Attention loss, where the CTC loss is back-propagated through the projection layer, while the Attention loss is back-propagated through the decoder. Refer to [18] for details.

3.2. Data pre-processing

The provided datasets of the challenge include the faces of the target speakers, so a pre-processing pipeline was designed to extract the lip region. Note that the pipeline was equally applied to both the training data (CN-CVS) and the development/evaluation data (CNVSRC-Single and CNVSRC-Multi).

Initially, we utilized RetinaFace [21] to detect the facial regions in each frame. Subsequently, FAN [22] was employed to extract facial landmarks, with which each detected face was aligned with a mean reference face. Finally, we extracted the lip region through the alignment for each video frame, which served as the input to the visual frontend. The modelling units are subword tokens, and the tokenizer was trained using the SentencePiece [23] tool in the unigram mode. During the training of this tokenizer, only the text data from CN-CVS was utilized. Subsequently, we employed this tokenizer to process the CN-CVS, CNVSRC-Single, and CNVSRC-Multi datasets, obtaining the token sequences used to train the recognition models.

3.3. Training strategy

We followed a two-step process to build the baseline systems. Firstly, we performed pre-training with the CN-CVS dataset. Subsequently, the development set was split into an 'adaptation set' and a 'validation set', with a ratio of 8:1 for the single-speaker dataset and 3:1 for the multi-speaker dataset. The adaptation set was used to fine-tune the pre-trained model, while the validation set was used to select the appropriate checkpoint. The training process was summarized in Table 2, and the details are as follows.

The initial training phase (P1) selected CN-CVS videos with a duration of less than 4 seconds to train an 'easy model'. The maximum number of training epochs was 75, and early stopping was triggered if the model's performance on the validation set started to drop. Once the training stopped, we selected the top 10 models based on their accuracy on the validation set, averaged their parameters to obtain the P1 model. Next, a full pre-training phase (P2) was evoked using the complete CN-CVS dataset, and the training was conducted for 75 epochs. Note that a warmup stage of 5 epochs was designed, by which the learning rate was gradually increased from 0 to 0.001. The average of the models of the last 10 epochs was used as the P2 model, i.e., the pre-trained model.

The fine-tuning step started from the P2 model and ran 80 epochs, including 2 epochs of warmup. This process is the same for the models trained for the single-speaker task and the multi-speaker task, but there are indeed some differences. Besides the learning rate (see Table 2), the most notable difference is that for the single-speaker model, we randomized the parameters of the classification layer of the P2 model, to provide sufficient space for the adaptation with the large amount of single-speaker data. The average of the models of the last 5 epochs was used as the final model, for both the single-speaker task and the multi-speaker task.

3.4. Performance

The performance of the baseline models was evaluated on the respective validation set and evaluation set for both the single-speaker task and multi-speaker task, using the TorchMetrics tool[2]. The CER results are as shown in Table 3.

Table 3: *Performance of the baseline systems.*

	Character Error Rate	
Task	T1: Single-speaker VSR	T2: Multi-speaker VSR
Valid	48.57%	58.77%
Eval	48.60%	58.37%

4. CNVSRC 2023 Report

4.1. Leaderboard

CNVSRC 2023 received 10 valid submissions from 6 teams. Most teams chose to submit their results to the single-speaker task, suggesting that single-speaker VSR is more suitable for the current stage of technical development, and multi-speaker VSR is over-challenging. The leaderboard is reported in Table 4.

Overall, T237 achieved the best performance in 3/4 of the tasks & tracks, and their results outperformed the baseline systems by a large margin. Their proposed system consists of a ResNet-3D visual frontend, an E-Branchformer encoder [24], and a Transformer decoder. The Chinese characters were used

[2]https://lightning.ai/docs/torchmetrics/stable/

Table 4: *Leaderboard of CNVSRC2023. Team ID and CER are reported.*

Task	T1: Single-speaker VSR				T2: Multi-speaker VSR			
Track	Fixed Track		Open Track		Fixed Track		Open Track	
Baseline	48.60%		48.60%		58.37%		58.37%	
Rank 1	T237	34.76%	T237	34.76%	T244	53.68%	T237	41.05%
Rank 2	T266	38.09%			T267	54.56%	T244	53.68%
Rank 3	T290	39.47%						
Rank 4	T238	40.52%						
Rank 5	T267	41.62%						

as the modelling units, and multiple data augmentation methods, including speed perturbation, random rotation, and horizontal flipping, were applied during training. Additionally, a ROVER-based system fusion [25] was performed during the inference procedure.

4.2. Technical summary

We summarize here the promising techniques demonstrated by the results of the submissions, highlighting the most effective methods in **bold font**.

4.2.1. Data pre-processing

Many teams adhered to the baseline system for data preprocessing. A notable observation is that T237 extracted lip regions of varying sizes and resolutions as inputs to their model and discovered that **larger regions (lip + mouth around)** yielded clear performance improvement.

4.2.2. Data augmentation

The participants extensively used data augmentation techniques, including random erase, random crop, random flip, and adaptive time masking. Notably, **speed perturbation** and **generative data augmentation** were reported to yield unexceptionally remarkable results. Speed perturbation adjusts the speed of the original videos by a factor ranging from 0.9 to 1.1. T237 and T238 observed notable enhancements in model performance (4.86% relative improvement) by applying speed perturbation. Generative data augmentation involves generating speech-driven lip videos, hence producing extra video-text training pairs. T238 utilized facial images from CNVSRC-Single and speech from CN-CVS and CNVSRC-Single to create an extra set of video-text pairs and reported a relative CER reduction of 6.98%.

4.2.3. Model structure

Most of the participating teams adopted the model architecture of the baseline systems, though some teams chose a more complicated backbone to pursue better performance. For instance, T237 achieved superior results using a **ResNet3D** structure. Moreover, T237 employed two advanced encoder structures: Branchformer [26] and **E-Branchformer** [24]. All these advanced structures lead to notable performance improvements. T266 introduced an **inner CTC residual module** [27, 28, 29] that resides in the Conformer block of the encoder. This module back-propagated a CTC loss through the shallow layers thus facilitating more effective parameter updates for the shallow layers of the model. Furthermore, taking inspiration from [30, 31], T266 utilized a **bi-transformer** structure to construct the decoder. This modification enhances the model's ability to capture contextual information from both the past and future segments.

4.2.4. Modeling units

Most participating teams used **Chinese characters as the modelling units**, and showed better performance than the subword tokens used by the baseline systems. In addition to subword tokens, T238 used **phonemes as supplementary modelling units** and designed a separate decoder for phoneme recognition. This approach achieved performance improvement, for which a hypothesis is that phonemes contain less semantic variation and thus are more closely related to lip movement, which may stabilize the training, especially in the early stage.

4.2.5. Cross-modality modeling

Some teams designed various approaches to **leverage the cross-modal dependency**. T290 invented an ASR-VSR joint system that forces the video representations to predict not only the text labels but also the speech representations produced by the middle layer of the ASR system. Following the same inspiration, T244 trained an audio-visual recognition system where the ASR and VSR have their respective independent encoders and decoders, and an AVSR decoder is constructed on top of the fused ASR and VSR features.

4.2.6. Decoding strategy

Several teams integrated RNN-based or Transformer-based language models to enhance the decoder, and mild performance improvement was reported. Moreover, system fusion was widely employed by teams to improve the performance of their systems.

5. Conclusion

This paper comprehensively details the inaugural Chinese Continuous Visual Speech Recognition Challenge (CNVSRC 2023). A key motivation of the challenge is to investigate the performance bound of VSR under the present data resource, e.g., 300 hours of training data from 2,557 speakers. The overall results suggest that the performance is far from satisfactory, even for the single-speaker scenario where about 100 hours of video is available for one person. The poor performance is certainly attributed to the lack of data, but whether it is related to the special linguistic properties of Chinese, e.g., the ubiquitous homophones, is unknown.

Based on these technical reports from participants, we have summarized the key techniques that might be crucial in constructing Chinese VSR systems. The most effective methods, ordered by their merit in terms of CER reduction: **Chinese characters as modelling units**, **rich data augmentation**, **fully 3D-CNN visual frontend**, **cross-modality modelling**, **system fusion**. Leveraging the technical insights provided by the participants, we have established a cutting-edge benchmark for Chinese LVC-VSR. We aspire that these resources will strengthen the burgeoning field of LVC-VSR research.

6. References

[1] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6319–6323.

[2] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2017, pp. 87–103.

[3] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2019, pp. 1–8.

[4] P. Ma, Y. Wang, J. Shen, S. Petridis, and M. Pantic, "Lip-reading with densely connected temporal convolutional networks," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 2857–2866.

[5] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8717–8727, 2018.

[6] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–11, 2018.

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

[8] J. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *INTERSPEECH*. ISCA, 2018.

[9] A. J. Goldschen, O. N. Garcia, and E. D. Petajan, "Continuous automatic speech recognition by lipreading," in *Motion-Based Recognition*. Springer, 1997, pp. 321–343.

[10] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2304–2308.

[11] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "Lipnet: End-to-end sentence-level lipreading," *arXiv preprint arXiv:1611.01599*, 2016.

[12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *International Conference on Machine learning (ICML)*, 2006, pp. 369–376.

[13] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[14] S. Jeon, A. Elsharkawy, and M. S. Kim, "Lipreading architecture based on multiple convolutional neural networks for sentence-level visual speech recognition," *Sensors*, vol. 22, no. 1, p. 72, 2021.

[15] B. Shillingford, Y. Assael, M. W. Hoffman, T. Paine, C. Hughes, U. Prabhu, H. Liao, H. Sak, K. Rao, L. Bennett *et al.*, "Large-scale visual speech recognition," *arXiv preprint arXiv:1807.05162*, 2018.

[16] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7613–7617.

[17] P. Ma, S. Petridis, and M. Pantic, "Visual speech recognition for multiple languages in the wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

[18] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[19] C. Chen, D. Wang, and T. F. Zheng, "Cn-cvs: A mandarin audio-visual dataset for large vocabulary continuous visual to speech synthesis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[20] Z. Gao, S. Zhang, I. McLoughlin, and Z. Yan, "Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition," in *INTERSPEECH*. ISCA, 2022.

[21] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *arXiv preprint arXiv:1905.00641*, 2019.

[22] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.

[23] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 66–71.

[24] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 84–91.

[25] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*. IEEE, 1997, pp. 347–354.

[26] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding," in *International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 17 627–17 643.

[27] J. Nozaki and T. Komatsu, "Relaxing the Conditional Independence Assumption of CTC-Based ASR by Conditioning on Intermediate Predictions," in *INTERSPEECH*, 2021, pp. 3735–3739.

[28] J. Lee and S. Watanabe, "Intermediate loss regularization for ctc-based speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6224–6228.

[29] M. Burchi and R. Timofte, "Audio-visual efficient conformer for robust speech recognition," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2258–2267.

[30] D. Wu, B. Zhang, C. Yang, Z. Peng, W. Xia, X. Chen, and X. Lei, "U2++: Unified two-pass bidirectional end-to-end model for speech recognition," *arXiv preprint arXiv:2106.05642*, 2021.

[31] B. Zhang, D. Wu, Z. Peng, X. Song, Z. Yao, H. Lv, L. Xie, C. Yang, F. Pan, and J. Niu, "Wenet 2.0: More productive end-to-end speech recognition toolkit," *arXiv preprint arXiv:2203.15455*, 2022.