CN-CVS: A MANDARIN AUDIO-VISUAL DATASET FOR LARGE VOCABULARY CONTINUOUS VISUAL TO SPEECH SYNTHESIS

Chen Chen[1,2], *Dong Wang*[1,*], *Thomas Fang Zheng*[1,*]

[1]Center for Speech and Language Technologies, BNRist, Tsinghua University, China
[2]Department of Computer Science and Technology, Tsinghua University, China

ABSTRACT

Research on Video to Speech Synthesis (VTS) surges recently and the focus is gradually shifting from small-vocabulary short-phrase VTS to large-vocabulary continuous VTS (LVC-VTS). A large-scale dataset with sufficient speakers and utterances is a prerequisite for such research, and the database is certainly language dependent.

In this paper, we introduce CN-CVS, a large-scale Mandarin continuous visual-speech dataset, to support LVC-VTS research. The dataset contains about 200k utterances from more than 2500 individuals, amounting to more than 300 hours of visual-speech data. We built a state-of-the-art VTS model with the new dataset and conducted preliminary studies. Our results show that models that achieve good performance on small vocabulary tasks may perform very poor on CN-CVS, indicating that continuous VTS is indeed a challenging task, and the main challenge comes from the unconstrained vocabulary. The dataset and baseline code can be downloaded for free from `http://cncvs.cslt.org`.

Index Terms— Mandarin, video to speech synthesis, audio-visual dataset

1. INTRODUCTION

Human perception of the world is multimodal. In conversations, for example, speech is the major modality, but people also use visual cues to assist message transmission, especially in noisy scenarios. Motivated by this observation, attempts to recover audio information from visual modality have been developed for years. For instance, in Audio-Visual Speech Enhancement (AVSE) [1, 2], visual information is used to help remove noise or non-target speech. In Audio-Visual Speech Recognition (AVSR) [3, 4, 5, 6], visual features are combined with acoustic features to improve the performance of speech-to-text transcription. In Visual Speech Recognition (VSR) [7, 8, 9], the goal is even more ambitious that tries to discover what is being spoken solely from visual information, mostly from lip movement.

Prof. Dong Wang (wangdong99@mails.tsinghua.edu.cn) and Prof. Thomas Fang Zheng (fzheng@tsinghua.edu.cn) are the corresponding authors. The work was supported by National Science Foundation of China (NSFC) with No. 62171250. Special thanks to Prof. Lantian Li from BUTP for his valuable suggestions.

Recently, Video to Speech Synthesis (VTS) is surging as a new attractive task [10, 11, 12, 13, 14, 15]. Compared to VSR, the main goal of VTS is restoring the corresponding *speech signal* from lip movement, rather than the *text in words*. An advantage of VTS compared to VSR is that more information is retained in the speech signal so that humans can make further judgments by listening. Another advantage is that VTS does not require word-annotated training data, which reduces the cost of system construction and makes it applicable to any language.

Most of existing VTS work focused on small-vocabulary tasks [14, 15, 16, 17], mostly using the GRID dataset [18]. This dataset contains utterances composed of 6 words, e.g., *place blue at F 9 now*. The size of the vocabulary is limited (51 words in total), and the angle of the camera is frontal. There were also attempts on large-vocabulary continuous VTS (LVC-VTS) [15, 16] using the TCD-TIMIT dataset [19]. More recently, some research started to investigate LVC-VTS with 'in-the-wild' complexity, e.g., unlimited vocabulary and varied camera positions [17, 20]. The LRS2 and LRS3 datasets [4] have been used for that purpose. The transition from small to large vocabulary tasks is much like the status of automatic speech recognition in the 1990s.

Despite the promising progress, we found that almost all the present datasets are in English. As for Chinese, the only two public VTS datasets are CAS-VSR-W1k [21] and CMLR [22]. The former contains only individual words, and the latter involves limited speakers. The lack of data resources not only places hindrances on building Chinese VTS models, but also prevents research on cross-lingual and multi-lingual methods. In this paper, we publish CN-CVS, a large-scale continuous visual-speech dataset in Mandarin Chinese, designed particularly for LVC-VTS research. It involves two parts: the News part that was collected from TV news with frontal cameras; the Speech part that was collected from speech shows with flexible cameras. In total, the dataset contains 200k utterances from more than 2500 individuals, amounting to more than 300 hours of visual-speech data. Table 1 compares CN-CVS and other datasets.

The rest of the paper is organized as follows. Section 2 introduces the principles considered during the design of CN-CVS, and Section 3 introduces details of the data. Section

Table 1. Datasets for VTS research

Name	Language	Type	Source	Transcript	Voc.	Speakers	Utterances	Duration	Camera angle	Speaker label
GRID [18]	English	Grammar patterns	Read	Human	51	33	33000	27h	0°	Yes
TCD-TIMIT [19]	English	Sentence	Read	Human	5954	62	6913	≈20h	$0^\circ, 30^\circ$	Yes
Lip2Wav [12]	English	Sentence	Lecture	No	≈5k/spk	5	-	≈120h	Natural	Yes
LRW [23]	English	Word	BBC	ASR, OCR	500	-	≈539000	173h	Natural	No
LRS [3]	English	Sentence	BBC	ASR, OCR	≈17k	-	118116	75.5h	Natural	No
LRS2 [4]	English	Sentence	BBC	ASR, OCR	≈60k	-	≈145000	224.5h	Natural	No
LRS3 [4]	English	Sentence	TED	ASR, OCR	≈70k	≈10000	≈165000	475h	Natural	No
CAS-VSR-W1k [21]	Mandarin	Word	TV show	Human	1k	>2000	718018	≈140h	Natural	No
CMLR [22]	Mandarin	Sentence	TV news	ASR	3517	11	102076	≈88h	0°	Yes
CN-CVS/News (Ours)	Mandarin	Sentence	TV news	No	-	28	13016	34.6h	0°	Yes
CN-CVS/Speech (Ours)	Mandarin	Sentence	Speech	No	-	2529	193245	273.4h	Natural	Yes

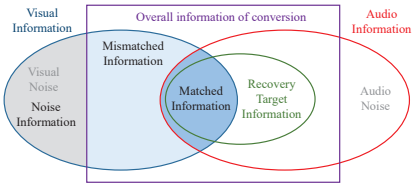

Fig. 1. Challenge of VTS from an information perspective.

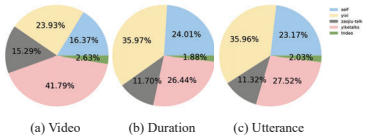

Fig. 2. Statics about data from different sources.

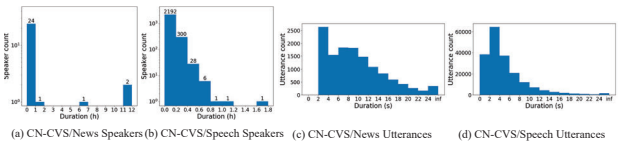

Fig. 3. Distribution of duration per speaker or utterance.

4 presents some experiments conducted with CN-CVS and other datasets, and Section 5 concludes the whole paper.

2. DESIGN PRINCIPLE

The VTS task can be analyzed from an information perspective, as shown in Fig. 1. Briefly, the visual stream provides three types of information: (1) Matched information (e.g., lip movement) that is important for restoring the desired speech; (2) Mismatched information (e.g., facial expression) that is useful for the conversation, but cannot be used in speech restoration; (3) Noise information (e.g., camera angle) that is useless, even harmful to the conversation. From the speech side, part of the information required for speech restoration can be obtained from the visual stream, while other information is audio specific so cannot be predicted from videos, for example, pitch pattern and speaker property. Some review or survey articles [24, 25] on Lip Reading task have similar content on the difficulty of VSR task, but the content about information is different from ours.

Based on the above analysis, we conclude the following principles to design data for LVC-VTS: (1) As vocabulary is large and the utterances are continuous and long, *matched information* is more deficient. We, therefore, need more data to cover the content variation; (2) Since diverse speakers produce much *mismatched information*, more speakers are required to learn speaker diversity if we want to train multi-speaker or speaker-independent systems; (3) Camera positions and other interfering factors may produce additional *noise information*. To make the model robust, training data should involve these variations. (4) Acoustic noise should always be avoided. The CN-CVS dataset was designed following the above principles and holds the following features.

- **Complex content**. The data was collected from TV news and Internet speech shows, to make sure the content is sufficiently complex and diverse.
- **Lots of speakers**. We collected data of 2500 speakers, ranging in age from teenagers to seniors. Most of them are not professional announcers, to ensure diversity in speaking styles.
- **Complex environmental factors**. The data collected involves videos recorded with diverse cameras, and in the same video, the angle and distance of the camera may change.
- **Moderate noise**. Most of the audio samples are clean and have no negative impact on human perception. In terms of video, the speaker's lips are visible most of the time, but in some frames may be occluded by hands or microphones.

3. DETAILS OF CN-CVS

3.1. Data profile

The CN-CVS dataset involves two parts: CN-CVS/News which consists of TV news, and CN-CVS/Speech which consists of speech shows from online media. The former is relatively constrained in speaking style and camera setting, while the latter involves more real-life complexity. Table 1 presents the data volume. Fig. 2 and Fig. 3 present some statistics in source and duration. For more detailed statistics please refer to `http://cncvs.cslt.org`.

3.2. Collection pipeline

To collect the data, we first searched for some video candidates using a web crawler, and then conducted preliminary human screening and speaker labeling. Secondly, similar to previous audio-visual datasets[4, 21, 22], we applied an automated pipeline to capture audio-visual synchronized segments. Finally, a human check was applied to ensure the quality. The automated pipeline is as follows.

- **Step 1. Shot change detection**. Given that VTS requires continuous lip movement, shot change should be avoided in a single training sample. We used the ffmpeg tool[1] to detect shot changes and cut the video into clips at the detected positions.
- **Step 2. Face detection**. We used the dlib toolkit[2] to perform face detection for each clip and removed clips with no faces or multiple faces. To speed up the process, face detection was conducted on frames evenly sampled from each clip. We also compared the detected faces in different frames and removed the clips where the sizes of the faces are clearly different. This helped delete clips derived from an incorrect shot change detection.
- **Step 3. Clip segmentation**. We used the pydub toolkit[3] to detect the silent frames and split a clip into short segmentations according to the detection result. Each segment roughly corresponds to a sentence.
- **Step 4. Face extraction**. The dlib toolkit was used to track the face. Firstly, for each frame, we obtained the face region and computed the center point of the region. Secondly, the trace of the center points was smoothed by a Savitzky-Golay filter. Thirdly, the maximum value of the width and height of the face areas in all the frames was used to construct a window, by which the face images in all the frames were extracted.
- **Step 5. Synchronization detection**. The final step was to ensure the visual and speech streams were synchronized. We used a pre-trained SyncNet[26] model[4] to perform the synchronization detection and removed segments with a low degree of synchronization.

4. EXPERIMENTS

We built a VCA-GAN model [13] to demonstrate the value of the CN-CVS dataset. VCA-GAN obtained competitive performance on GRID and can be seen as one of the state-of-the-art models. The source code from the authors of the original paper was adopted to reproduce the performance on GRID[5]. The same code was used to build models with CN-CVS, and the only difference is that we simply extracted the lip region using a landmark detection model[6], without further face alignment.

4.1. Evaluation metrics

Following the convention of previous research, we used Perceptual Evaluation of Speech Quality (PESQ), Short Term Objective Intelligibility (STOI), and Extended Short Term Objective Intelligibility (ESTOI) as the evaluation metrics. Additionally, an ASR system was also used to test the intelligibility of the recovered speech, for which Word Error Rate (WER) was used as the metric. The ASR models used for the GRID system were trained using the same data used to train the GRID VTS model. For the CN-CVS system, we employed an ASR model from PaddleSpeech[7].

We also conducted a human listening test in which 26 Chinese students participated. They were asked to rate the intelligibility of the generated speech, given the ground truth transcription. The test was conducted with randomly selected 300 samples for each people and the averaged MOS value was reported.

4.2. Basic results

In the first experiment, we built VCA-GAN models with GRID, CN-CVS/News, and CN-CVS/Speech respectively. The test was performed under three settings: single speaker, multi-seen speakers, and multi-unseen speakers.

Under the single speaker setting, a model was trained for each speaker and tested on the same speaker. Under the multi-seen speaker setting, data from several speakers were pooled to train a single model, and the model was tested with each speaker in the training set. Under the multi-unseen speaker setting, a single model was trained with data from multiple speakers, and then was tested with speakers absent in the training set. The data profile for the three settings is presented in Table 2. Note that the profile of the GRID experiment follows the convention of previous research [14, 13].

The results are shown in Table 3. An immediate observation is that the performance on CN-CVS is clearly worse than the results on GRID, indicating that LVC-VTS on CN-CVS is more challenging than small-vocabulary VTS on GRID. Most significantly, the performance on CN-CVS/News under the single-speaker setting is also very poor, although the condition is not substantially more complex than GRID and the data volume is even larger. This clearly indicates that the primary challenge for LVC-VTS is the vocabulary size, rather than the speaker or device variety.

The second notable observation is that comparing the results of the single speaker setting and the multi-seen speaker

[1] https://ffmpeg.org/
[2] http://dlib.net/
[3] https://pypi.org/project/pydub/
[4] https://github.com/joonson/syncnet_python
[5] https://github.com/ms-dot-k/Visual-Context-Attentional-GAN
[6] https://github.com/midasklr/98-FaceLandmarks
[7] https://github.com/PaddlePaddle/PaddleSpeech

Table 2. Data splitting in the basic experiments.

Test	Dataset	Speaker		
		train	test	validation
single	GRID	s1, s2, s4, s29	s1, s2, s4, s29	s1, s2, s4, s29
	News	n002∼n005	n002∼n005	n002∼n005
	Speech	s00001∼s00005	s00001∼s00005	s00001∼s00005
multi-seen	GRID	all speakers	all speakers	all speakers
	News	n001∼n005	n001∼n005	n001∼n005
	Speech	s00001∼s01434	s00001∼s01434	s00001∼s01434
multi-unseen	GRID	s1, s3, s5, s6, s7, s8, s10, s12, s14, s16, s17, s22, s24, s26, s28, s32	s2, s4, s11, s13, s15, s18, s19, s25, s31, s33	s9, s20, s23, s27, s29, s30, s34
	News	n001∼n006, n009, n010	n007, n008, n014∼n028	n011, n012, n013
	Speech	s00001∼s01250	50% of s01251∼s02529	other 50% of s01251∼s02529

setting, the latter is better on and only on CN-CVS/News. This seems to indicate that if the data variation is limited (e.g., no variation in camera angles) and the data volume of each speaker is large, multi-speaker training helps improve performance. Note that on CN-CVS/Speech, the multi-speaker training reduced the performance, indicating that noise information might be a major hindrance to training speaker-independent systems.

Table 3. Basic results with GRID and CN-CVS.

Test	Dataset	Objective Metrics				MOS
		PESQ	STOI	ESTOI	WER %	
Single	GRID	1.933	0.679	0.492	12.699	4.695
	News	1.496	0.383	0.180	92.263	1.918
	Speech	1.297	0.358	0.110	97.611	1.449
Multi-Seen	GRID	1.908	0.694	0.493	7.542	4.664
	News	1.494	0.416	0.199	81.445	2.431
	Speech	1.303	0.311	0.052	99.433	1.102
Multi-Unseen	GRID	1.407	0.579	0.297	29.627	4.454
	News	1.368	0.348	0.128	96.341	1.608
	Speech	1.294	0.296	0.046	99.284	1.080

4.3. Additional results

We provide two additional experiments to demonstrate the value of CN-CVS as a large-scale data resource.

4.3.1. Value in large-scale training

Table 4. Results with CAS-VSR-W1k and CN-CVS/Speech

Test set	Evaluation Metrics	Train set	
		CN-CVS/Speech	CAS-VSR-W1k
CN-CVS/Speech	PESQ	1.294	**1.353**
	STOI	**0.296**	0.259
	ESTOI	**0.046**	0.004
	MOS	**1.080**	1.030
CAS-VSR-W1k	PESQ	1.579	**1.592**
	STOI	**0.284**	0.278
	ESTOI	**0.053**	0.041
	MOS	**1.101**	1.067

In this experiment, we trained two models with CN-CVS/Speech and CAS-VSR-W1k [21], a word-level Chinese VTS dataset, respectively, and tested the models on the test set of each other. Since there are no speaker labels in CAS-VSR-W1k, we only conducted experiments with the multi-unseen speaker setting. The results are shown in Table 4. A significant observation is that on the CAS-VSR-W1k test set, the CN-CVS model achieved better results than the CAS-VSR-W1k model in terms of all the metrics except PESQ, although the words in the CAS-VSR-W1k test set have multiple instances in the CAS-VSR-W1k training data.[8] This better performance is clearly attributed to the complex and large-scale data in CN-CVS, and the result suggests that LVC-VTS training is powerful and can be simply adopted to solve word-level VTS tasks.

4.3.2. Value in pre-training

In this experiment, CN-CVS/Speech was used as an additional data resource to pre-train VTS models. Four speakers in the CMLR dataset [22] were used to perform the test, under the single speaker setting. For each speaker we trained two models using 1-hour data: one was trained from scratch and the other was initialized with a model pre-trained using CN-CVS/Speech under the multi-unseen speaker setting.

Table 5 shows the results. It can be seen that pre-training with CN-CVS makes the model easier to achieve good performance when the speaker-specific data is inadequate. In general, fine-tuning the pre-trained model for 50 epochs can obtain better results than the model trained from scratch for 500 epochs.

Table 5. Results on CMLR w/w.o. CN-CVS pre-training

Spk	Pre-train dataset	Fine-tune Epochs	Evaluation Metrics			
			PESQ	STOI	ESTOI	MOS
s2	-	500	1.359	0.321	0.152	1.994
	CN-CVS/Speech	50	**1.373**	**0.361**	**0.167**	**2.092**
s3	-	500	1.282	0.344	0.176	**2.000**
	CN-CVS/Speech	50	**1.295**	**0.365**	**0.180**	1.895
s7	-	500	1.224	0.328	0.116	1.414
	CN-CVS/Speech	50	**1.253**	**0.353**	**0.144**	**1.722**
s9	-	500	1.320	0.323	0.096	1.318
	CN-CVS/Speech	50	**1.348**	**0.337**	**0.132**	**1.611**

5. CONCLUSION

We presented an open-source audio-visual dataset CN-CVS to support research on large vocabulary continuous VTS (LVC-VTS). The dataset contains 200k sentences from over 2500 speakers and involves sufficient diversity in camera settings and spontaneous speaking styles. We demonstrated the value of the dataset with a start-of-the-art VTS model. Two major conclusions of the experiments are: (1) LVC-VTS is very challenging, mostly due to the large vocabulary; (2) CN-CVS is a valuable data resource and can be used reliably in LVC-VTS research.

[8] Note that from the results in Table 3, PESQ seems not an appropriate metric and often shows a different trend compared to other metrics, in particular MOS.

6. REFERENCES

[1] Zhiqi Kang et al., "The Impact of Removing Head Movements on Audio-Visual Speech Enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7302–7306, IEEE.

[2] Xinmeng Xu et al., "VSEGAN: Visual Speech Enhancement Generative Adversarial Network," in *ICASSP 2022*, Singapore, Singapore, May 2022, pp. 7308–7311, IEEE.

[3] Joon Son Chung et al., "Lip Reading Sentences in the Wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, July 2017, pp. 3444–3453, IEEE.

[4] Triantafyllos Afouras et al., "Deep Audio-visual Speech Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[5] Takaki Makino et al., "Recurrent Neural Network Transducer for Audio-Visual Speech Recognition," Nov. 2019, arXiv:1911.04890 [cs, eess].

[6] Bowen Shi et al., "Robust Self-Supervised Audio-Visual Speech Recognition," Mar. 2022, arXiv:2201.01763 [cs, eess].

[7] Yannis M. Assael et al., "LipNet: End-to-End Sentence-level Lipreading," Dec. 2016, arXiv:1611.01599 [cs].

[8] Brendan Shillingford et al., "Large-Scale Visual Speech Recognition," Oct. 2018, arXiv:1807.05162 [cs].

[9] K. R. Prajwal et al., "Sub-word Level Lip Reading With Visual Attention," Dec. 2021, arXiv:2110.07603 [cs].

[10] Ariel Ephrat et al., "Improved Speech Reconstruction from Silent Video," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 455–462, ISSN: 2473-9944.

[11] Konstantinos Vougioukas et al., "Video-Driven Speech Reconstruction Using Generative Adversarial Networks," in *Interspeech 2019*. pp. 4125–4129, ISCA.

[12] K. R. Prajwal et al., "Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis," May 2020, arXiv:2005.08209 [cs, eess].

[13] Minsu Kim et al., "Lip to Speech Synthesis with Visual Context Attentional GAN," in *Advances in Neural Information Processing Systems*. 2021, vol. 34, pp. 2758–2770, Curran Associates, Inc.

[14] Dan Oneaţă et al., "Speaker disentanglement in video-to-speech conversion," in *2021 29th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 46–50, ISSN: 2076-1465.

[15] Nasir Saleem et al., "E2E-V2SResNet: Deep residual convolutional neural networks for end-to-end video driven speech synthesis," *Image and Vision Computing*, vol. 119, pp. 104389, Mar. 2022.

[16] Rodrigo Mira et al., "End-to-End Video-to-Speech Synthesis Using Generative Adversarial Networks," *IEEE Transactions on Cybernetics*, pp. 1–13, 2022, Conference Name: IEEE Transactions on Cybernetics.

[17] Rodrigo Mira et al., "SVTS: Scalable Video-to-Speech Synthesis," May 2022, arXiv:2205.02058 [cs, eess].

[18] Martin Cooke et al., "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, Nov. 2006.

[19] Naomi Harte et al., "TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015, Conference Name: IEEE Transactions on Multimedia.

[20] Leyuan Qu et al., "LipSound2: Self-Supervised Pre-Training for Lip-to-Speech Reconstruction and Lip Reading," Dec. 2021, arXiv:2112.04748 [cs, eess].

[21] Shuang Yang et al., "LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild," Apr. 2019, arXiv:1810.06990 [cs].

[22] Ya Zhao et al., "A Cascade Sequence-to-Sequence Model for Chinese Mandarin Lip Reading," Nov. 2019, arXiv:1908.04917 [cs].

[23] Joon Son Chung et al., "Lip Reading in the Wild," vol. 10112, pp. 87–103. Springer International Publishing, Cham, 2017.

[24] Adriana Fernandez-Lopez, "Survey on automatic lip-reading in the era of deep learning," *Image and Vision Computing*, p. 20, 2018.

[25] Marzieh Oghbaie, Arian Sabaghi, Kooshan Hashemifard, and Mohammad Akbari, "Advances and Challenges in Deep Lip Reading," Oct. 2021, arXiv:2110.07879 [cs].

[26] Joon Son Chung et al., "Out of Time: Automated Lip Sync in the Wild," in *Computer Vision – ACCV 2016 Workshops*, Chu-Song Chen, Jiwen Lu, and Kai-Kuang Ma, Eds., vol. 10117, pp. 251–263. Cham, 2017.