– Supplementary Material –

VoiceCraft-Dub: Automated Video Dubbing with Neural Codec Language Models

In this supplementary material, we provide details on the dataset construction pipeline, additional results and analysis, implementation details, metrics, and human evaluations that were not included in the main paper.

Contents

A. Dataset and code

For reproducibility, we plan to open-source our training and inference code, along with the curated dataset and its annotations, upon acceptance.

B. Advantage of the autoregressive method

We chose the autoregressive (AR) method due to its proven efficacy in speech generation, outperforming non-autoregressive (NAR) methods in zero-shot TTS, as demonstrated by VoiceCraft [15]. We utilize this pretrained AR decoder, which inherently ensures high-quality speech and effectively preserves voice characteristics via in-context learning. Note that our goal is to demonstrate the potential of AR modeling for video dubbing, not to exclude NAR methods. Achieving state-of-the-art results in video dubbing justifies the use of the AR approach, though we acknowledge that NAR methods could also be effective.

Visatronic [9] is a concurrent work that also uses the AR method for this task. Visatronic relies on a speaker embedding model, which is known to yield low similarity in TTS [23], and entirely omits evaluations of speaker similarity. In contrast, our model supports robust instant voice cloning via in-context learning, effectively preserving speaker similarity even for unseen speakers. In addition, our approach significantly improves lip-sync and content accuracy (both comparable to ground truth) by directly fusing lip and audio tokens at each timestep, whereas Visatronic prepends all modalities in the token space.

	Train	Test	Total
Number of total video clips	67,549	216	67,765
Number of speakers	6,530	33	6,563
Average utterances per speaker	10.34	6.55	10.33
Average duration (seconds)	4.58	3.39	4.57

Table S1. **Statistics of our curated CelebV-Dub dataset.**

C. Data curation pipeline for CelebV-Dub

We introduce the CelebV-Dub dataset, consisting of expressive video clips specifically suitable for automated video dubbing tasks. Despite the abundance of existing talking-video datasets [1, 6, 19, 20, 24], our goal is to curate in-the-wild videos that capture natural yet expressive speech. Such videos are effective for training and testing automated video dubbing models, which require synthesizing not only neutral but also expressive speech synchronized with facial cues. Our curated dataset comprises multiple speakers and utterances, each accompanied by a corresponding transcript. The dataset statistics are summarized in Table S1.

Video collection. We initially collect videos from existing sources, including CelebV-HQ [28] and CelebV-Text [27]. These datasets originate from diverse sources, such as vlogs, dramas, and influencer videos, providing expressive, in-the-wild content across various scenarios. However, the provided metadata from these datasets varies in length—from single utterances to longer sequences—and contains substantial noise, such as non-active speakers and occluded faces. Therefore, we design a data curation pipeline to collect suitable videos specifically for automated video dubbing.

Detecting English and labeling pseudo-transcripts. For each video in the existing sources, we use WhisperX [2] to detect the language and generate pseudo-transcripts automatically. In constructing this dataset, we retain only videos identified as containing English speech, discarding all others.

Trimming videos into utterances. WhisperX provides timestamps at the word and sentence levels, enabling precise video segmentation. Given the variability in utterance lengths of the original videos, we unify the dataset by trimming each video clip to contain a single utterance, utilizing the timestamps provided by WhisperX.

Frontal face verification. The trimmed videos occasionally contain faces that are not oriented toward the front, preventing the models from learning distinct facial movements corresponding to speech. To address this, we measure yaw and pitch angles using Mediapipe [11] and remove clips with abrupt head movements or large yaw and pitch angles, which indicate side-facing poses.

Active speaker detection. Training videos for automated video dubbing require facial movements synchronized with speech. To ensure this synchronization, we apply TalkNet [21], a model that employs audio-visual cross-attention to identify active speakers. We set conservative thresholds to minimize false positives, ensuring that only videos clearly containing active speakers are retained. Clips that do not meet these thresholds are discarded.

Background music suppression. Background music in audio tracks can disturb clear speech signals necessary for effective model training. We employ Spleeter [10] to detect and suppress background music where present, thereby preserving the clarity of speech signals.

Speaker classification. Finally, we classify utterances by speaker identity. Initially, videos extracted from the same source video are grouped together. However, since we cannot guarantee that all clips from a single video contain the same speaker, we apply an off-the-shelf speaker recognition model [3] to measure pairwise speaker similarity. Clips within the same source video are then re-clustered based on these similarity scores, using a defined threshold to determine speaker identity clusters.

D. Analysis on lip-synchronzation metric

As lip synchronization accuracy (LSE-D) measured by SyncNet [5] has been shown to be unstable in several studies [12, 25, 26], we investigate whether these findings align within our dataset and cases. Specifically, we analyze the correlation between LSE-D and human evaluation of lip synchronization on the same samples. For this analysis, we use SyncNet to measure the LSE-D for a synthesized speech sample and its corresponding video, while five human evaluators assess the lip synchronization of the same sample using the Mean Opinion Score (MOS). LSE-D is a distance metric, where lower values indicate better lip synchronization, while MOS uses a 1-5 rating scale, with higher values indicating better quality. To facilitate comparison, we reverse the LSE-D score (by taking the negative) to align it with the MOS scale. The MOS for each sample is averaged from the ratings of five human evaluators. A total of 50 samples are used in this analysis.

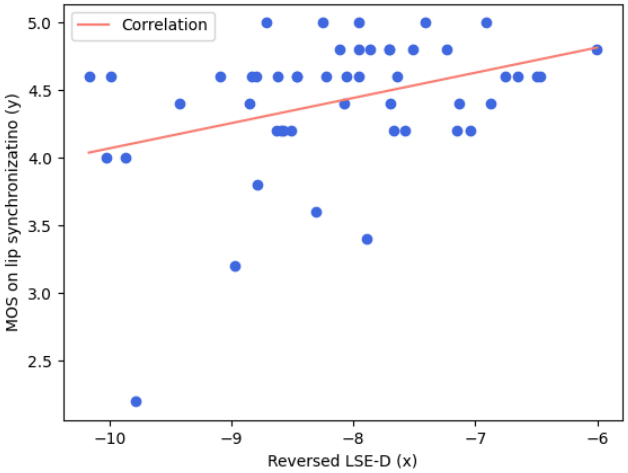


Figure S1. **Correlation between human evaluation and lip-sync objective metric.** We visualize the scatter plot showing the relationship between the objective lip-sync metric (LSE-D) and the subjective Mean Opinion Score (MOS) on lip-sync from human evaluation. We observe a weak correlation between the two, with a correlation coefficient of 0.36, indicating that LSE-D should be used as a reference rather than a definitive metric.

Figure S1 shows a scatter plot with reversed LSE-D on the x-axis and MOS of lip synchronization on the y-axis. Interestingly, we observe a weak correlation between the objective and subjective metrics, with a correlation coefficient of 0.36. Furthermore, even when the LSE-D scores are relatively high (indicating poor lip-sync according to the objective metric), ranging from 7 to 10, human ratings mostly remain above 4, which is considered a relatively high score on the MOS scale. The average LSE-D for the 50 samples in this analysis is 8.11, while the average MOS is 4.42. This suggests that, despite relatively poor LSE-D scores, humans perceive the lip synchronization as sufficiently accurate.

Given the weak correlation between LSE-D and human evaluation, we conclude that human evaluation is the most accurate metric for validating lip synchronization performance. While LSE-D remains a useful objective metric for evaluating lip synchronization, as this analysis shows, it should not be considered definitive; rather, it serves better as a reference metric when human evaluation is limited.

A vs. B	Naturalness			Expressiveness			Lip synchronization		
	A wins (%)	Neutral	B wins (%)	A wins (%)	Neutral	B wins (%)	A wins (%)	Neutral	B wins (%)
Ours vs. HPMDubbing [7]	**99.2**	0.4	0.4	**96.4**	0.4	3.2	**88.8**	3.2	8.0
Ours vs. StyleDubber [8]	**98.0**	0.4	1.6	**99.2**	0.4	0.4	**91.6**	6.0	2.4
Ground-Truth vs. HPMDubbing [7]	99.6	0.0	0.4	98.8	0.4	0.8	89.8	7.1	3.1
Ground-Truth vs. Ours	58.4	24.4	17.2	57.2	26.4	16.4	44.0	40.0	16.0

Table S2. **A/B testing results on CelebV-Dub.** We report the preferences (%) between A and B across various aspects of synthesized speech. In rows 1 and 2, our model is significantly preferred by humans over existing methods. Comparing rows 3 and 4, HPMDubbing is significantly less preferred compared to the ground truth, while our model, highlighted with a gray background , is preferred more. Surprisingly, over 41.6% of the time, our model is perceived as equally good as or better than the ground truth.

Method	WER (↓)	LSE-D (↓)	LSE-C (↑)	spkSIM (↑)	UTMOS (↑)	DNSMOS (↑)	MCD (↓)	F0 (↓)	Energy (↓)	emoSIM (↑)
Ground-Truth	5.96	7.28	7.35	-	3.10	3.44		-	-	-
Zero-shot TTS [15]	2.97	12.52	2.08	0.279	3.02	3.54	-	-	-	0.733
HPMDubbing [7]	17.36	**6.98**	**7.65**	0.176	2.62	3.10	9.11	129.30	4.87	0.759
StyleDubber [8]	16.06	11.38	3.18	0.248	2.38	3.00	8.48	136.16	3.75	0.790
Ours (lip-only)	8.91	7.62	6.97	**0.321**	**3.55**	3.51	**7.17**	88.12	**2.91**	0.769
Ours (lip & face)	**8.11**	7.71	6.88	0.312	**3.55**	**3.52**	7.34	**86.01**	2.94	**0.791**

Table S3. **Generalization results on Voxceleb2.** We highlight the best results in **bold** and underline the second best among all the methods.

E. Additional results and analysis

E.1. Human evaluation

We present the A/B testing results on our curated CelebV-Dub dataset in Table S2. Similar to the main paper, our model is significantly preferred by humans over existing methods, with over 98% preference for naturalness, 96% for expressiveness, and 88% for lip synchronization. Comparing rows 3 and 4, we observe that, in most cases, the ground truth is preferred over HPMDubbing, while our model is rated as good as or better than the ground truth over 41.6% of the time across all metrics. These results indicate that, on the CelebV-Dub dataset, which contains expressive content, our model synthesizes speech that is both temporally and semantically aligned with the target video while being sufficiently expressive, leading to high human preference.

It is worth noting that our model performs favorably against the ground truth and outperforms existing methods in lip synchronization during A/B testing, even though our lip synchronization metrics in Sec. 4.3 of the main paper show lower performance than HPMDubbing [7]. These results in Sec. 4.3 may be due to the instability of SyncNet, a limitation discussed in Sec. D of the supplementary material and related works [12, 25, 26]. Therefore, human evaluation should be given more weight to validate the lip-sync accuracy of the synthesized speech.

E.2. Generalization results

We incorporate a subset of the VoxCeleb2 [6] dataset to evaluate the generalization performance of our approach. Both our proposed models and the comparison models are trained on the LRS3 dataset [1] and tested on the VoxCeleb2 subset. For this experiment, we select 200 samples from the VoxCeleb2 test split and use the Whisper [16] large model to extract pseudo ground-truth text for each sample. As summarized in Table S3, our methods outperform the other approaches across most of the metrics, particularly demonstrating a substantial improvement in WER and automatic MOS evaluations. Both our models perform lower than HPMDubbing on LSE-D/C. However, this can still be considered satisfactory, as the LSE-D scores are better than those in Sec. D (average LSE-D of 8.11), which achieved an average MOS of 4.42, indicating sufficiently good lip synchronization. Interestingly, our model variant using both lip and face input shows a significant improvement in emotional similarity (emoSIM) compared to the lip-only variant, highlighting the advantage of combining both inputs for expressive speech synthesis.

E.3. Qualitative results

We visually compare the mel-spectrogram samples synthesized by prior methods [7, 8] and our proposed approach, along with those from the ground-truth recordings in Fig. S2. As shown in the results, HPMDubbing often produces incorrect speech, failing to convey accurate content. While StyleDubber performs better than HPMDubbing in terms of content accuracy, its synthesized signal is often blurry, indicating considerable noise, and it frequently exhibits time misalignment (see columns 2 and 3). In contrast, the samples generated by our model accurately convey content with clear and distinct mel frequencies, closely matching the ground-truth mel-spectrograms. These results demonstrate the superiority of our model over existing methods in producing accurate, time-aligned, and high-quality speech.

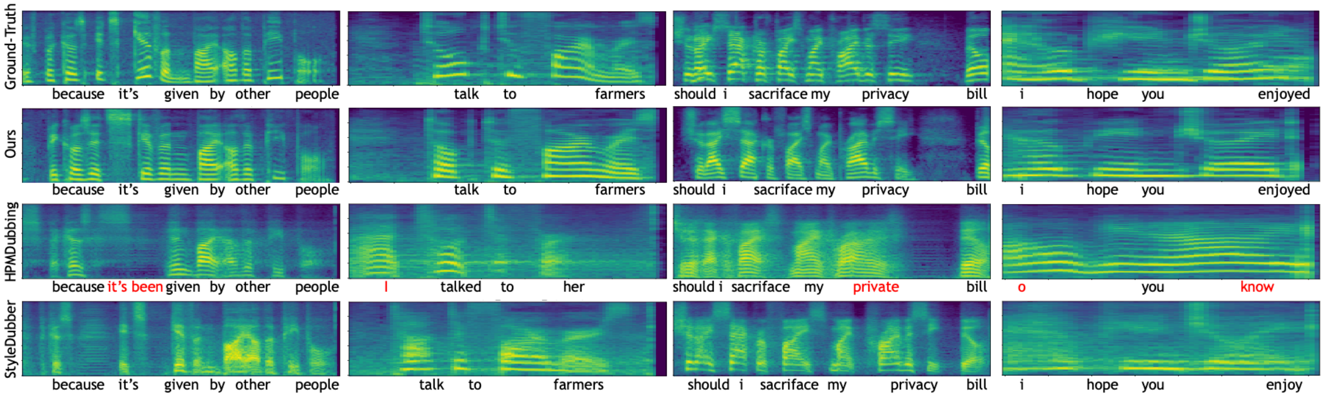


Figure S2. **Qualitative results.** We compare mel-spectrogram visualizations from ground-truth recordings, our model, and existing methods [7, 8] on the LRS3 (columns 1–2) and our CelebV-Dub (columns 3–4) datasets. The texts below each mel-spectrogram represent time-aligned speech extracted using Whisper [16], with red text indicating incorrectly synthesized speech.

F. Details on the objective metrics

WER. Word Error Rate (WER) is a widely used metric in the speech-to-text domain. Since the output of automated video dubbing is speech, we rely on an off-the-shelf ASR model to extract text from the synthesized speech and measure the WER. Specifically, we use the Whisper [16] medium.en model to extract text from the synthesized speech and measure WER.

LSE-D and LSE-C. To measure lip synchronization accuracy, we assess the audio-visual synchronization between lip movements and speech. We use SyncNet [5], which has learned representations for aligning lip movements with corresponding speech snippets. Two metrics are measured using SyncNet: Lip Sync Error - Distance (LSE-D) and Lip Sync Error - Confidence (LSE-C). LSE-D measures the Euclidean distance between the audio and visual embeddings extracted by SyncNet, where lower values indicate better audio-visual synchronization. LSE-C is a probability-based confidence metric derived from the embeddings' distances, with higher values indicating higher confidence in synchronization.

Speaker similarity (spkSIM). We use WavLM-TDNN [3] to measure speaker similarity. As we prompt the source speech from the same speaker as the target speech but a different utterance, we assume the model synthesizes the speech in the target speaker's voice. After synthesizing the speech, we measure the cosine similarity between the synthesized speech features and the ground-truth target speech using the WavLM-TDNN embedding space.

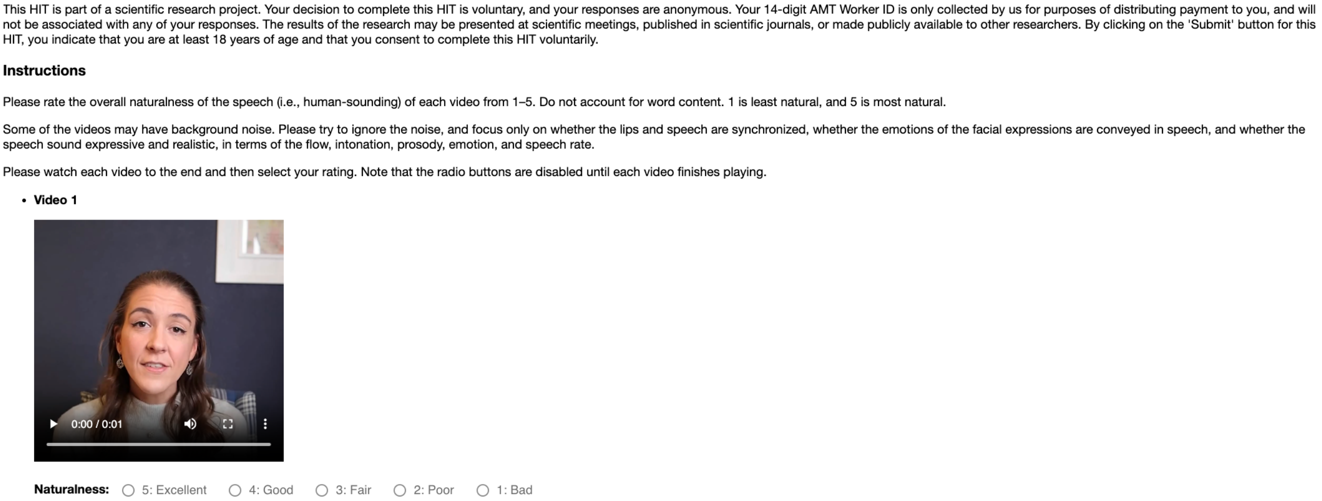
Emotion similarity (emoSIM). We measure the expressiveness of the synthesized speech by evaluating the emotion similarity between the synthesized speech and the ground truth. We use the Emotion2Vec [22] model to compute the cosine similarity between the synthesized speech and the ground truth in its embedding space.

DNSMOS and UTMOS. Deep Noise Suppression MOS (DNSMOS) [17] and Universal TTS MOS (UTMOS) [18] are used to objectively evaluate speech quality by approximating subjective human ratings (Mean Opinion Score, MOS). DNSMOS is designed to assess the quality of speech processed by noise suppression algorithms, measuring clarity, naturalness, background noise quality, and overall quality. Similarly, UTMOS focuses on evaluating the quality of synthesized speech, particularly by assessing naturalness, intelligibility, prosody, and expressiveness.

MCD, energy, and F0. We measure several low-level metrics: Mel-Cepstral Distortion (MCD), F0 distance (F0), and energy distance (Energy). MCD is used to measure the intelligibility of speech, while F0 and Energy are more closely correlated with prosody similarity between the synthesized speech and the ground truth. We follow the implementation of these metrics in [15]. MCD measures the difference in Mel Frequency Cepstrum Coefficients (MFCC) between the generated and ground-truth speech, using a 13-order MFCC and the pymcd package for measurement. For F0 measurement, we use the pYIN algorithm [13], implemented in librosa [14], with minimal and maximal frequencies set to 80Hz and 600Hz, respectively. The energy distance is computed using the root mean square of the magnitude of the spectrogram, extracted via the short-time Fourier transform with a window length of 640 and a hop size of 160.

G. Additional implementation details

Training setup. We introduce two variants of our model in the main paper: one with lip-only input and the other with both lip and face input. We observe that the lip-only variant yields favorable results compared to existing work and ground truth. For training the latter model, we find that starting with the lip-only model and zero-initializing the AV fusion layer for full face input leads to stable training. Furthermore, when training on the CelebV-Dub dataset, we



This HIT is part of a scientific research project. Your decision to complete this HIT is voluntary, and your responses are anonymous. Your 14-digit AMT Worker ID is only collected by us for purposes of distributing payment to you, and will not be associated with any of your responses. The results of the research may be presented at scientific meetings, published in scientific journals, or made publicly available to other researchers. By clicking on the 'Submit' button for this HIT, you indicate that you are at least 18 years of age and that you consent to complete this HIT voluntarily.

Instructions

Please rate the overall naturalness of the speech (i.e., human-sounding) of each video from 1–5. Do not account for word content. 1 is least natural, and 5 is most natural.

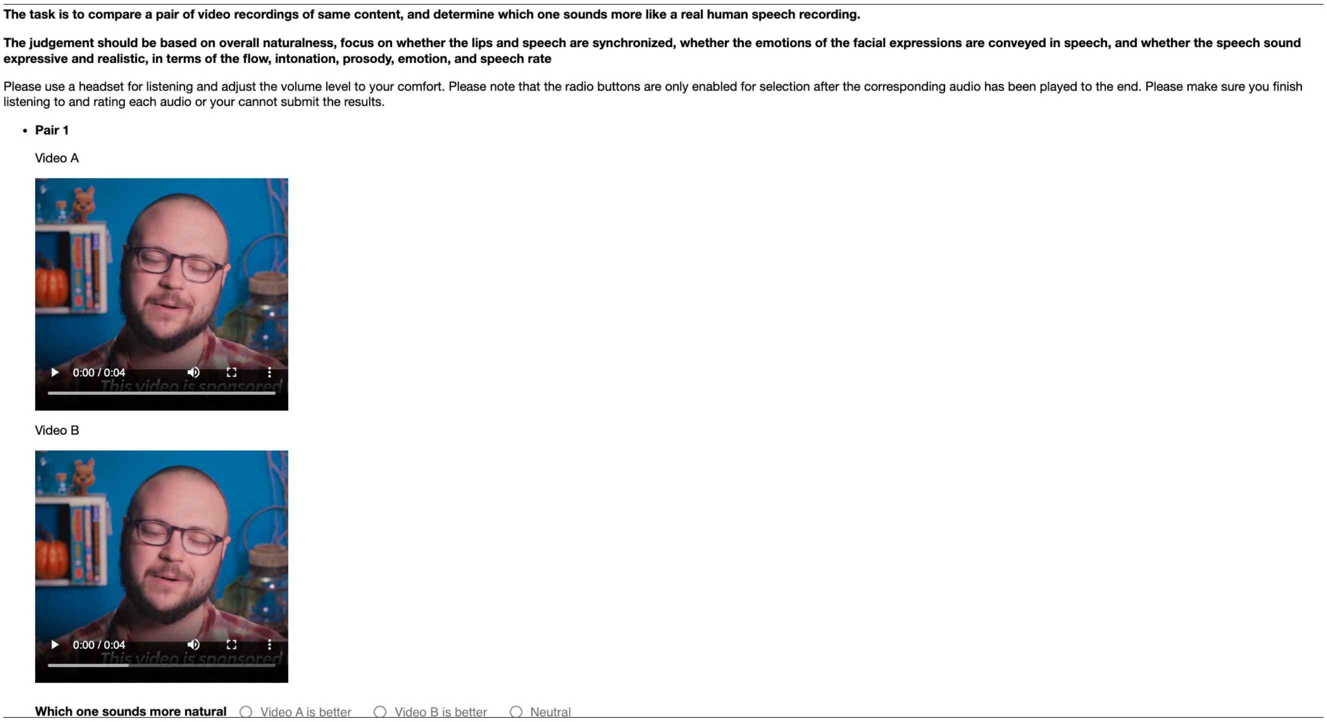
Some of the videos may have background noise. Please try to ignore the noise, and focus only on whether the lips and speech are synchronized, whether the emotions of the facial expressions are conveyed in speech, and whether the speech sound expressive and realistic, in terms of the flow, intonation, prosody, emotion, and speech rate.

Please watch each video to the end and then select your rating. Note that the radio buttons are disabled until each video finishes playing.

- **Video 1**

Naturalness: ○ 5: Excellent ○ 4: Good ○ 3: Fair ○ 2: Poor ○ 1: Bad

Figure S3. **Instruction and sample for AMT human listening test on overall naturalness of speech.**



The task is to compare a pair of video recordings of same content, and determine which one sounds more like a real human speech recording.

The judgement should be based on overall naturalness, focus on whether the lips and speech are synchronized, whether the emotions of the facial expressions are conveyed in speech, and whether the speech sound expressive and realistic, in terms of the flow, intonation, prosody, emotion, and speech rate

Please use a headset for listening and adjust the volume level to your comfort. Please note that the radio buttons are only enabled for selection after the corresponding audio has been played to the end. Please make sure you finish listening to and rating each audio or your cannot submit the results.

- **Pair 1**

Video A

Video B

Which one sounds more natural ○ Video A is better ○ Video B is better ○ Neutral

Figure S4. **Instruction and sample for AMT human listening test for A/B testing on overall naturalness of speech.**

initialize the model with a version pretrained on the LRS3 dataset. We apply the same training setup to existing methods' [7, 8] training to ensure a fair comparison.

Inference setup. Although our model synthesizes high quality, natural, and lip-synced speech, autoregressive generation may sometimes result in inaccurate output. Therefore, as mentioned in Sec.4.1 of the paper, we design a sorting strategy similar to VALL-E 2 [4]. Given ten synthesized speech samples, ${\mathbf{Y}_{\text{tgt, i}}}_{i=1}^{10}$, we sort them using content accuracy (WER) and lip synchronization accuracy (LSE-D) to select the optimal sample. We denote the WER and LSE-D values for each sample as $\mathbf{Y}^{\text{WER}}_{\text{tgt, i}}$ and $\mathbf{Y}^{\text{LSE-D}}_{\text{tgt, i}}$, respectively. Specifically, we first sort the samples according to LSE-D if the WER is below 5%, and otherwise, we sort them based on WER, where lower values are preferred. This sorting method is defined as:

$$\mathbf{Y}_{\text{tgt, best}} = \arg\min_{\mathbf{Y}_{\text{tgt, i}}}([\min(\mathbf{Y}^{\text{WER}}_{\text{tgt, i}}, 0.05), \mathbf{Y}^{\text{LSE-D}}_{\text{tgt, i}}]). \quad (1)$$

This sorting strategy is also applied to existing methods across all evaluations to ensure a fair comparison.

H. Details on the human evaluation

Amazon Mechanical Turk (AMT) is used to conduct human listening tests. We select 100 audio samples from the LRS3 test set and 50 audio samples from the CelebV-Dub test set, totaling 400 samples for LRS3 and 200 samples for CelebV-Dub, with samples from the three models and the ground truth. For the mean opinion score (MOS), we design an extensive evaluation based on various criteria: naturalness, intelligibility, expressiveness, lip synchronization, and speaker similarity. We use a 5-point Likert scale, where 1 represents "poor" and 5 represents "excellent." In the A/B testing, we present two samples to a Turker and ask them to judge which one sounds better in terms of naturalness, expressiveness, or lip synchronization, allowing them to choose either sample as better or neutral. For each sample or comparison, 5 ratings are obtained from different Turkers. We also compute the 95% confidence interval for MOS. In the MOS test, 43 Turkers participated in the LRS3 listening test, and 25 Turkers participated in the CelebV-Dub listening test. For A/B testing, 34 Turkers participated in the LRS3 test, and 23 Turkers participated in the CelebV-Dub test. Please refer to Fig. S3 and Fig. S4 for sample instructions.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 1, 3

[2] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2023. 2

[3] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 2022. 2, 4

[4] Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*, 2024. 5

[5] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, 2017. 2, 4

[6] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Conference of the International Speech Communication Association (INTERSPEECH)*, 2018. 1, 3

[7] Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang, Ming-Hsuan Yang, and Qingming Huang. Learning to dub movies via hierarchical prosody models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4, 5

[8] Gaoxiang Cong, Yuankai Qi, Liang Li, Amin Beheshti, Zhedong Zhang, Anton Hengel, Ming-Hsuan Yang, Chenggang Yan, and Qingming Huang. Styledubber: Towards multi-scale style learning for movie dubbing. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. 3, 4, 5

[9] Akshita Gupta, Tatiana Likhomanenko, Karren Dai Yang, Richard He Bai, Zakaria Aldeneh, and Navdeep Jaitly. Visatronic: A multimodal decoder-only model for speech synthesis. *arXiv preprint arXiv:2411.17690*, 2024. 1

[10] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 2020. 2

[11] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2

[12] Junxian Ma, Shiwen Wang, Jian Yang, Junyi Hu, Jian Liang, Guosheng Lin, Kai Li, Yu Meng, et al. Sayanything: Audio-driven lip synchronization with conditional video diffusion. *arXiv preprint arXiv:2502.11515*, 2025. 2, 3

[13] Matthias Mauch and Simon Dixon. pyin: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014. 4

[14] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. *SciPy*, 2015. 4

[15] Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. *arXiv preprint arXiv:2403.16973*, 2024. 1, 3, 4

[16] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023. 3, 4

[17] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022. 4

[18] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*, 2022. 4

[19] Kim Sung-Bin, Lee Chae-Yeon, Gihun Son, Oh Hyun-Bin, Janghoon Ju, Suekyeong Nam, and Tae-Hyun Oh. Multitalk: Enhancing 3d talking head generation across languages with multilingual video dataset. In *Conference of the International*

Speech Communication Association (INTERSPEECH), 2024. 1

[20] Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, and Tae-Hyun Oh. Laughtalk: Expressive 3d talking head generation with laughter. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2024. 1

[21] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *ACM International Conference on Multimedia (MM)*, 2021. 2

[22] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021. 4

[23] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023. 1

[24] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[25] Dogucan Yaman, Fevziye Irem Eyiokur, Leonard Bärmann, Seymanur Akti, Hazım Kemal Ekenel, and Alexander Waibel. Audio-visual speech representation expert for enhanced talking face video generation and evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[26] Dogucan Yaman, Fevziye Irem Eyiokur, Leonard Bärmann, Hazım Kemal Ekenel, and Alexander Waibel. Audio-driven talking face generation with stabilized synchronization loss. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[27] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[28] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European Conference on Computer Vision (ECCV)*, 2022. 1