V2SFlow: Video-to-Speech Generation with Speech Decomposition and Rectified Flow

Jeongsoo Choi[1*], *Ji-Hoon Kim*[1*], *Jinyu Li*[2], *Joon Son Chung*[1], *Shujie Liu*[2]
[1]Korea Advanced Institute of Science and Technology, [2]Mircrosoft
{jeongsoo.choi, jh.kim, joonson}@kaist.ac.kr, {jinyli, shujliu}@microsoft.com

***Abstract*—In this paper, we introduce V2SFlow, a novel Video-to-Speech (V2S) framework designed to generate natural and intelligible speech directly from silent talking face videos. While recent V2S systems have shown promising results on constrained datasets with limited speakers and vocabularies, their performance often degrades on real-world, unconstrained datasets due to the inherent variability and complexity of speech signals. To address these challenges, we decompose the speech signal into manageable subspaces (content, pitch, and speaker information), each representing distinct speech attributes, and predict them directly from the visual input. To generate coherent and realistic speech from these predicted attributes, we employ a rectified flow matching decoder built on a Transformer architecture, which models efficient probabilistic pathways from random noise to the target speech distribution. Extensive experiments demonstrate that V2SFlow significantly outperforms state-of-the-art methods, even surpassing the naturalness of ground truth utterances.**

***Index Terms*—video-to-speech, speech decomposition, rectified flow matching, diffusion transformer**

I. Introduction

Video-to-Speech (V2S) synthesis is an emerging research area aiming to generate natural-sounding speech solely from lip movements captured in silent video. This technology holds the potential to reconstruct human-like speech when the auditory signal is missing, opening up a variety of practical applications, such as enabling silent communication in secure environments and providing assistive technologies for individuals with aphonia. The advent of deep learning has revolutionized V2S synthesis by leveraging synchronized speech and lip movement sequences from video data for training, eliminating the need for additional annotations such as text transcriptions or speaker labels. This self-supervised nature significantly enhances the efficiency and scalability of V2S systems, paving the way for broader real-world applications.

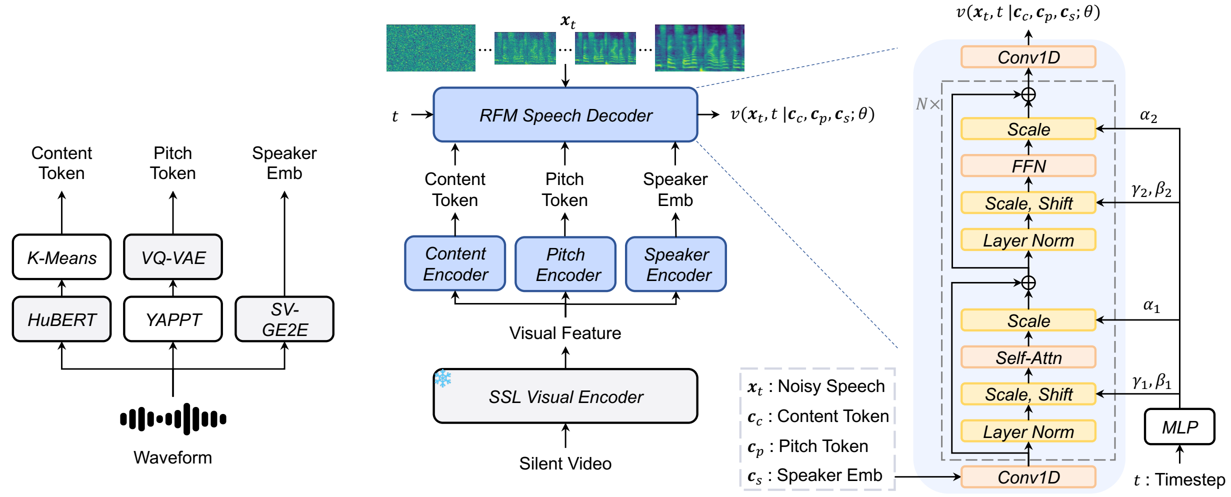
Traditional V2S systems are typically trained on relatively small datasets with substantial constraints, such as controlled environments and a limited number of speakers [1]–[3]. While these systems demonstrate promising results on such constrained datasets, their performance deteriorates when applied to real-world datasets with a larger number of speakers and a broader vocabulary. The primary challenge in achieving high-quality synthesis in real-world scenarios arises from the complex nature of speech signals, which encompass various acoustic components, including linguistic content and speaker characteristics. These components interact in nuanced ways, making it difficult for V2S systems to generalize effectively beyond the specific conditions of their training data.

To improve the generation quality in real-world scenarios, numerous studies have been conducted, broadly fall into two categories. The first focuses on modeling the inherent variability in speech to address the ambiguity between lip motions and the corresponding speech. To mitigate this ambiguity, several approaches incorporate text labels [4], self-supervised speech units [5]–[8], or extra lip-reading networks [9], [10]. Other studies aim to clarify speaker identity by injecting speaker representations derived from either reference audio [11] or the input video itself [12]. The second research line seeks to better capture the complex dynamics of speech by leveraging advanced model structures and training algorithms, such as normalizing flows [13], generative adversarial networks [14]–[16], and diffusion models [10], [12]. Despite these advancements, current V2S systems still face challenges due to the inherent complexities of speech. Synthesized speech often suffers from artifacts, highlighting the need for more powerful architecture or training strategy to produce more natural outputs.

In this paper, we propose V2SFlow, a novel framework designed to generate natural and intelligible speech from silent video. To address the inherent ambiguity between lip movements and corresponding speech, we decompose speech into three fundamental and manageable subspaces: content, pitch, and speaker characteristics [17], [18]. For each speech sample, we extract content tokens, pitch tokens, and speaker embedding as representative features, which serve as prediction targets for training a video-to-speech model. Our model includes three specialized encoders, built upon a Self-Supervised Learning (SSL) visual encoder, to predict each of these decomposed attributes. With the predicted attributes, we employ a Rectified Flow Matching (RFM)-based decoder with Transformer backbone to reconstruct authentic speech. This approach combines the strengths of RFM [19] and Diffusion Transformer (DiT) [20], enabling high-quality speech generation with a small number of sampling steps. Our extensive evaluations demonstrate the effectiveness of V2SFlow, even achieving superior naturalness compared to the ground truth speech. Audio samples can be found on our demo page[1].

[*] These authors contributed equally to this work. This work was supported by the National Research Foundation of Korea grant (MSIT, RS-2023-00212845) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant (MSIT, RS-2024-00457882, National AI Research Lab Project) funded by the Korean government.

[1] https://mm.kaist.ac.kr/projects/V2SFlow



(a) Speech Decomposition (b) Overall Architecture (c) RFM Speech Decoder

Fig. 1. Overall framework of V2SFlow. In subfigure (a), SV-GE2E refers to the speaker verification model trained with GE2E loss. In subfigure (b), $\boldsymbol{x}_t$ denotes the intermediate noisy mel-spectrogram at timestep t, and v represents the corresponding vector field. The parameters of the SSL video encoder are not updated during training. In subfigure (c), α, β, and γ represent the scale and shift parameters derived from the timestep.

II. Method

A. Speech Decomposition

Considering the complexity inherent in speech, we factorize it into distinct speech attributes and separately estimate them from silent video. This simplifies the video-to-speech mapping, thereby facilitating more stable and effective training.

Content. Accurate linguistic content is crucial for generating intelligible speech, as it helps resolve ambiguities between lip movements and spoken words [21]. To extract this linguistic information without relying on additional text labels, we utilize the self-supervised speech model HuBERT [22]. In particular, we employ HuBERT tokens discretized by K-means algorithm, as these tokens are known to capture detailed linguistic information independent of paralinguistic cues [23]–[25].

Pitch. Pitch is a fundamental speech component, essential for conveying prosody and adding nuance beyond literal content. Building on recent works [17], [26], [27], we extract the pitch sequence from raw audio using `YAPPT` [28], and then normalize it for each sentence to obtain distinctive pitch variations while reducing speaker information. This extracted pitch sequence is processed by a pre-trained VQ-VAE [29], which encodes the sequence into a discrete latent space. The codebook from VQ-VAE provides quantized indices representing the pitch sequence, which we refer to as pitch tokens. These tokens serve as the representation of pitch in training our system, facilitating effective modeling and synthesis.

Speaker. We extract a global speaker embedding from a pre-trained speaker verification model optimized with GE2E loss [30]. This embedding captures speaker-specific characteristics across entire time dimension, helping the model reduce ambiguity caused by the varying traits of different speakers.

B. Speech Attribute Estimation

Given the decomposed speech attributes, we estimate each attribute directly from silent video based on recent findings that reveal unique associations between visual and acoustic features [31]–[33]. To do this, we utilize pre-trained AV-HuBERT (Large) [32] model, a powerful SSL audio-visual network capable of extracting both rich visual and linguistic information from the input video. Based on the visual features obtained from AV-HuBERT, we estimate each speech attribute through their respective encoders, as illustrated in Fig. 1(b).

We utilize Conformer [34] for the content, pitch, and speaker encoders, as it is well-known for effectively capturing rich and contextualized features. Given the decomposed content tokens $\boldsymbol{c}_c$ from the target speech, as mentioned earlier, the content encoder is trained to estimate $\boldsymbol{c}_c$ from SSL visual features, and optimized by a frame-level cross-entropy (CE) loss with a label smoothing parameter set to $\alpha = 0.1$:

$$\mathcal{L}_c = (1 - \alpha) CE(\boldsymbol{c}_c, \hat{\boldsymbol{c}}_c) + \alpha CE(\boldsymbol{u}, \hat{\boldsymbol{c}}_c), \tag{1}$$

where $\hat{\boldsymbol{c}}_c$ refers to the estimated content tokens, and $\boldsymbol{u}$ denotes a uniform distribution. Similar to the content encoder, the pitch encoder aims to predict the decomposed pitch tokens $\boldsymbol{c}_p$ and is trained with cross-entropy loss, also equipped with a label smoothing parameter α of 0.1.

$$\mathcal{L}_p = (1 - \alpha) CE(\boldsymbol{c}_p, \hat{\boldsymbol{c}}_p) + \alpha CE(\boldsymbol{u}, \hat{\boldsymbol{c}}_p), \tag{2}$$

where $\hat{\boldsymbol{c}}_p$ refers to the generated pitch tokens. The speaker encoder takes SSL visual features as inputs and generates a single global speaker embedding by averaging the output features along the temporal dimension. To train this speaker encoder, we apply a cosine similarity loss defined as:

$$\mathcal{L}_s = 1 - \cos(\boldsymbol{c}_s, \hat{\boldsymbol{c}}_s), \tag{3}$$

where $\boldsymbol{c}_s$ and $\hat{\boldsymbol{c}}_s$ represent the decomposed target and predicted speaker embeddings, respectively. Note that each encoder, including the subsequent speech decoder, is trained separately.

TABLE I
Video-to-speech synthesis performance comparisons on LRS3-TED and LRS2-BBC dataset. ↑: higher is better, ↓: lower is better.

Method	LRS3-TED					LRS2-BBC				
	UTMOS↑	WER↓	SECS↑	MAE$_{F0}$↓	LSE-C↑	UTMOS↑	WER↓	SECS↑	MAE$_{F0}$↓	LSE-C↑
Ground Truth	3.519	2.5	–	–	7.63	3.017	4.2	–	–	8.15
with speaker embedding from audio										
SVTS [11]	1.256	78.0	0.557	0.389	6.04	1.349	80.9	0.593	0.374	7.91
Intelligible [6]	2.657	29.8	0.761	0.265	**8.04**	2.294	38.1	0.701	0.278	8.23
V2SFlow-A	**3.624**	**28.5**	**0.851**	**0.245**	7.97	**3.393**	**35.2**	**0.819**	**0.263**	**8.28**
with speaker embedding from video										
DiffV2S [12]	2.989	39.2	0.627	0.290	7.28	2.877	51.9	0.568	0.306	7.45
LTBS [7]	2.428	79.7	0.607	0.289	7.84	2.319	86.4	0.534	0.306	7.74
V2SFlow-V	**3.780**	**28.5**	**0.664**	**0.251**	**8.09**	**3.648**	**35.6**	**0.581**	**0.275**	**8.39**

C. *Speech Decoder*

To reconstruct high-fidelity mel-spectrograms from the decomposed attributes, we introduce a Rectified Flow Matching (RFM)-based decoder with Transformer [35] backbone. RFM [19] estimates the vector field of the probability path from random noise to target data samples. Similar to diffusion-based models, it models the probability path between a tractable prior and target distributions. However, RFM aims to build straight paths between the prior distribution $\boldsymbol{x}_0 \sim p_0(\boldsymbol{x})$ and the target distribution $\boldsymbol{x}_1 \sim p_1(\boldsymbol{x})$, seeking to minimize the number of sampling steps. Given the condition $\boldsymbol{c}$ and an intermediate data sample $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1$ at timestep $t \in [0,1]$, RFM generative model θ is trained to estimate the vector field $u(\boldsymbol{x}_t, t|\boldsymbol{x}_1, \boldsymbol{c}) = \boldsymbol{x}_1 - \boldsymbol{x}_0$ with the RFM objective:

$$\mathcal{L}_{RFM} = ||v(\boldsymbol{x}_t, t|\boldsymbol{c};\theta) - (\boldsymbol{x}_1 - \boldsymbol{x}_0)||^2, \tag{4}$$

where $v(\boldsymbol{x}_t, t|\boldsymbol{c};\theta)$ is the estimated vector field.

In our case, we construct the condition $\boldsymbol{c}$ by concatenating all speech attributes along the channel dimension[2]. Inspired by DiT [20], we utilize Transformer-based backbone and adaLN-Zero block to condition $\boldsymbol{t}$, facilitating effective generation. Moreover, we amplify the conditional sampling trajectory by adopting Classifier-Free Guidance (CFG) [36]. During training, we randomly drop all condition with a probability of 0.1. During inference, the speech decoder iteratively refines the mel-spectrogram, guiding the sampling trajectory away from unconditional flows. We use an Euler solver with CFG: $\boldsymbol{x}_{t+\epsilon} = \boldsymbol{x}_t + \epsilon\{\gamma v(\boldsymbol{x}_t, t|\boldsymbol{c};\theta) + (1-\gamma)v(\boldsymbol{x}_t, t|\varnothing;\theta)\}$, where ϵ and γ denote the step size and guidance scale, respectively.

III. Experimental Settings

A. *Datasets*

We train our model on LRS3-TED [37], which contains video clips featuring thousands of speakers and over 50,000 words. To enable unseen speaker evaluation, we adopt the unseen train-test split from SVTS [11]. Furthermore, to evaluate the generalizability in diverse environments, we use LRS2-BBC [38] solely for inference. The video frame rate of both datasets is 25 fps, and the audio sample rate is 16 kHz.

B. *Evaluation Metrics*

We evaluate our method on each dataset using various metrics. UTMOS [39] estimates perceptual naturalness and Word Error Rate (WER) assesses the intelligibility of audio. To calculate the WER, we use a pre-trained speech recognition model [40] to transcribe the audio clips and compare them to the ground truth text labels. We measure Speaker Embedding Cosine Similarity (SECS), which assess the similarity between speaker representations[3] of both the synthesized and target speech. In addition, we obtain Mean Absolute Error of normalized F0 (MAE$_{F0}$) between the synthesized and target speech to assess the pitch accuracy, and also measure LSE-C [41] using the pre-trained SyncNet [42] to evaluate lip synchronization accuracy.

C. *Implementation Details*

We crop, convert to grayscale, and apply data augmentation to the input video clips, as described in [6]. To crop the video, we detect the face using RetinaFace [43] and extract facial landmarks using FAN [44]. Meanwhile, the audio is converted into 80-bin mel-spectrogram with a hop size of 160 and a window size of 640. The resulting mel-spectrogram is then normalized based on the maximum and minimum values from the training dataset. We stack 20ms of the mel-spectrogram to obtain 50Hz feature.

We follow the Conformer block design from previous works [6], [11] for our encoder architecture, and the speech decoder includes an 8-layer Transformer with a hidden dimension of 512 and 4 attention heads. Content, pitch, and speaker encoders are separately trained for 50k steps with a batch size of 144 seconds per GPU. We use 8 GPUs and adopt the same learning rate schedule as in [6]. The SSL visual encoder is kept frozen during training and the speech decoder is trained for 400k steps with a batch size of 384 seconds on a single GPU. We adopt logit-normal sampling [45] for sampling timestep t and the number of sampling steps is set to 30. To convert mel-spectrogram to audible waveform, we employ HiFi-GAN [46].

D. *Baseline Systems*

To investigate the impact of two different speaker embeddings—one derived from reference speech and the other

[2] We match the temporal lengths of speech attributes through linear interpolation. During training, the decomposed ground truth attributes are used, while during inference, the predicted values are utilized.

[3] We use `Resemblyzer` python library to extract speaker representations.

TABLE II
MOS RESULTS ON LRS3-TED DATASET.

Method	Naturalness↑	Intelligibility↑	Similarity↑
Ground Truth	4.42±0.12	4.88±0.06	4.92±0.05
SVTS [11]	1.03±0.03	1.49±0.12	1.23±0.09
Intelligible [6]	2.46±0.13	3.11±0.19	2.85±0.20
DiffV2S [12]	3.28±0.16	3.04±0.19	2.61±0.16
LTBS [7]	2.43±0.13	1.87±0.13	2.27±0.14
V2SFlow-A	4.38±0.10	3.90±0.17	**4.11±0.15**
V2SFlow-V	**4.60±0.10**	**3.91±0.17**	3.38±0.19

TABLE III
ABLATION STUDY OF SPEECH DECOMPOSITION ON LRS3-TED DATASET.

Method	UTMOS↑	WER↓	SECS↑	MAE$_{F0}$ ↓	LSE-C↑
V2SFlow-V	3.780	28.5	0.664	0.251	8.09
w/o $\mathbf{c}_c$	3.271	106.4	0.643	0.259	2.81
w/o $\mathbf{c}_p$	3.796	28.5	0.663	0.255	8.11
w/o $\mathbf{c}_s$	3.722	28.7	0.587	0.252	7.99
w/o Decomp.	3.534	30.4	0.623	0.262	8.38

predicted from silent video—we conduct experiments with two model variations: V2SFlow-A and V2SFlow-V. V2SFlow-A is compared to SVTS [11] and Intelligible [6], both of which use audio-driven speaker embeddings. V2SFlow-V is compared with two other baselines, DiffV2S [12] and LTBS [7], which leverage video-driven speaker embeddings.

IV. Experimental Results

A. Quality Comparison

We evaluate the quality of V2SFlow against recent systems using both objective and subjective metrics, with the results presented in Table I and Table II, respectively. In Table I, V2SFlow consistently outperforms existing methods by a large margin. Our method even achieves a superior UTMOS compared to the ground truth, indicating that the synthesized speech from our method is more natural than the ground truth speech. The lowest WER indicates that V2SFlow synthesizes authentic speech with plausible linguistic content, while the best results in SECS and MAE$_{F0}$ demonstrate that the synthesized speech closely resembles the ground truth in terms of voice quality and pitch variations.

In addition, we conduct subjective Mean Opinion Score (MOS) tests on the LRS3-TED dataset, where 30 domain experts rate the quality of 40 randomly selected audio clips on a scale from 1 to 5. As shown in Table II, our method significantly outperforms existing methods in terms of naturalness, intelligibility, and similarity. Consistent with the UTMOS results, our method achieves better naturalness than the ground truth speech. This implies that our method focuses solely on lip movement, resulting in clear speech that is independent of the background noise present in the ground truth speech.

B. Ablation studies

1) Speech Attributes: We explore the effect of decomposed speech attributes, with the results summarized in Table III. The findings demonstrate that each attribute makes an independent contribution to overall speech quality. Specifically, excluding $\boldsymbol{c}_c$ significantly degrades intelligibility and synchronization accuracy, as reflected by WER and LSE-C. The absence of $\boldsymbol{c}_p$ results in reduced pitch accuracy, as indicated by a higher MAE$_{F0}$. The speaker attribute $\boldsymbol{c}_s$ also plays a crucial role, with a noticeable drop in SECS when it is excluded. When we directly generate mel-spectrograms from visual features instead of estimating decomposed speech attributes, the overall quality is highly degraded, demonstrating the effectiveness of speech decomposition in V2S system.

TABLE IV
ABLATION STUDY OF SPEECH DECODER ON LRS3-TED DATASET.

Method	UTMOS↑	WER↓	SECS↑	MAE$_{F0}$↓	LSE-C↑
V2SFlow-V	3.780	28.5	0.664	0.251	8.09
w/ RFM → DDIM	3.596	28.9	0.663	0.252	7.87
w/ adaLN-Zero → Concat	3.625	29.1	0.662	0.251	7.83

TABLE V
ABLATION STUDY OF GUIDANCE SCALE ON LRS3-TED DATASET.

γ	UTMOS↑	WER↓	SECS↑	MAE$_{F0}$↓	LSE-C↑
1	3.365	29.0	0.659	0.250	7.67
1.5	3.722	28.7	0.665	0.251	8.02
2	3.780	28.5	0.664	0.251	8.09
4	3.541	28.7	0.655	0.251	8.03

2) Speech Decoder: Table IV presents the results of our ablation study on the speech decoder. In the second row, we replace RFM with the Denoising Diffusion Implicit Model (DDIM) [47], maintaining the same number of sampling steps (30). The results consistently demonstrate RFM's superiority over DDIM, with notable improvements in UTMOS, WER, SECS, and LSE-C metrics. Moreover, in the third row, we replace the adaLN-Zero conditioning with simple concatenation, as employed in DiffV2S [12]. This modification results in further performance degradation, particularly evident in the WER, SECS, and LSE-C scores, indicating the effectiveness of the adaLN-Zero conditioning method.

3) Guidance scale: To confirm the trade-off introduced by the guidance scale γ, we conduct a series of experiments, as shown in Table V. The first row demonstrates the clear advantage of using CFG, as $\gamma = 1$—which corresponds to not using CFG—results in the worst performance. We find that $\gamma = 2$ yields the best overall results, with only a slight trade-off in SECS and MAE$_{F0}$.

V. Conclusion

In this paper, we propose V2SFlow, which can effectively generate high-quality speech from silent videos. We focus on the intrinsic modelling complexity of speech, and address it by decomposing speech into fundamental speech attributes. Based on these attributes, we introduce a rectified flow-based speech decoder with Transformer architecture to generate high-fidelity speech with preserving acoustic and linguistic characteristics. Through extensive experiments, we have demonstrated the effectiveness of V2SFlow, and a comprehensive analysis further validates the reliability of our design.

References

[1] T. Le Cornu and B. Milner, “Generating intelligible audio speech from visual speech,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 25, no. 9, pp. 1751–1761, 2017. 1

[2] A. Ephrat and S. Peleg, “Vid2speech: speech reconstruction from silent video,” in *Proc. ICASSP*, 2017. 1

[3] Y. Kumar, R. Jain, K. M. Salik, R. R. Shah, Y. Yin, and R. Zimmermann, “Lipper: Synthesizing thy speech using multi-view lipreading,” in *Proc. AAAI*, 2019. 1

[4] M. Kim, J. Hong, and Y. M. Ro, “Lip-to-speech synthesis in the wild with multi-task learning,” in *Proc. ICASSP*, 2023. 1

[5] W.-N. Hsu, T. Remez, B. Shi, J. Donley, and Y. Adi, “Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration,” in *Proc. CVPR*, 2023. 1

[6] J. Choi, M. Kim, and Y. M. Ro, “Intelligible lip-to-speech synthesis with speech units,” in *Proc. Interspeech*, 2023. 1, 3, 4

[7] J.-H. Kim, J. Kim, and J. S. Chung, “Let there be sound: Reconstructing high quality speech from silent videos,” in *Proc. AAAI*, 2024. 1, 3, 4

[8] S. Lei, X. Cheng, M. Lyu, J. Hu, J. Tan, R. Liu, L. Xiong, T. Jin, X. Li, and Z. Zhao, “Uni-dubbing: Zero-shot speech synthesis from visual articulation,” in *Proc. ACL*, 2024. 1

[9] S. Hegde, R. Mukhopadhyay, C. Jawahar, and V. Namboodiri, “Towards accurate lip-to-speech synthesis in-the-wild,” in *Proc. ACM MM*, 2023. 1

[10] Y. Yemini, A. Shamsian, L. Bracha, S. Gannot, and E. Fetaya, “Lipvoicer: Generating speech from silent videos guided by lip reading,” in *Proc. ICLR*, 2024. 1

[11] R. Mira, A. Haliassos, S. Petridis, B. W. Schuller, and M. Pantic, “Svts: Scalable video-to-speech synthesis,” in *Proc. Interspeech*, 2022. 1, 3, 4

[12] J. Choi, J. Hong, and Y. M. Ro, “Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding,” in *Proc. ICCV*, 2023. 1, 3, 4

[13] J. He, Z. Zhao, Y. Ren, J. Liu, B. Huai, and N. Yuan, “Flow-based unconstrained lip to speech generation,” in *Proc. AAAI*, 2022. 1

[14] M. Kim, J. Hong, and Y. M. Ro, “Lip to speech synthesis with visual context attentional GAN,” in *NeurIPS*, 2021. 1

[15] R. Mira, K. Vougioukas, P. Ma, S. Petridis, B. W. Schuller, and M. Pantic, “End-to-end video-to-speech synthesis using generative adversarial networks,” *IEEE Transactions on Cybernetics*, vol. 53, no. 6, 2022. 1

[16] S. B. Hegde, K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, “Lip-to-speech synthesis for arbitrary speakers in the wild,” in *Proc. ACM MM*, 2022. 1

[17] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, “Speech resynthesis from discrete disentangled self-supervised representations,” in *Proc. Interspeech*, 2021. 1, 2

[18] H.-S. Choi, J. Lee, W. Kim, J. Lee, H. Heo, and K. Lee, “Neural analysis and synthesis: Reconstructing speech from self-supervised representations,” in *NeurIPS*, 2021. 1

[19] X. Liu, C. Gong, and Q. Liu, “Flow straight and fast: Learning to generate and transfer data with rectified flow,” in *Proc. ICLR*, 2023. 1, 3

[20] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *Proc. CVPR*, 2023. 1, 3

[21] M. Kim, J. H. Yeo, and Y. M. Ro, “Distinguishing homophenes using multi-head visual-audio memory for lip reading,” in *Proc. AAAI*, vol. 36, no. 1, 2022, pp. 1174–1182. 2

[22] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021. 2

[23] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, “On generative spoken language modeling from raw audio,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021. 2

[24] F. Kreuk, A. Polyak, J. Copet, E. Kharitonov, T.-A. Nguyen, M. Rivière, W.-N. Hsu, A. Mohamed, E. Dupoux, and Y. Adi, “Textless speech emotion conversion using discrete and decomposed representations,” in *Proc. EMNLP*, 2022. 2

[25] M. Kim, J. Choi, D. Kim, and Y. M. Ro, “Textless unit-to-unit training for many-to-many multilingual speech-to-speech translation,” *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 32, p. 3934–3946, 2024. 2

[26] A. Polyak, L. Wolf, Y. Adi, O. Kabeli, and Y. Taigman, “High fidelity speech regeneration with application to speech enhancement,” in *Proc. ICASSP*, 2021. 2

[27] H.-Y. Choi, S.-H. Lee, and S.-W. Lee, “Dddm-vc: Decoupled denoising diffusion models with disentangled representation and prior mixup for verified robust voice conversion,” in *Proc. AAAI*, 2024. 2

[28] K. Kasi and S. A. Zahorian, “Yet another algorithm for pitch tracking,” in *Proc. ICASSP*, 2002. 2

[29] A. Van Den Oord, O. Vinyals *et al.*, “Neural discrete representation learning,” in *NeurIPS*, 2017. 2

[30] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, “Generalized end-to-end loss for speaker verification,” in *Proc. ICASSP*, 2018. 2

[31] M. Hassid, M. T. Ramanovich, B. Shillingford, M. Wang, Y. Jia, and T. Remez, “More than words: In-the-wild visually-driven prosody for text-to-speech,” in *Proc. CVPR*, 2022. 2

[32] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, “Learning audio-visual speech representation by masked multimodal cluster prediction,” in *Proc. ICLR*, 2022. 2

[33] J. Lee, Y. Oh, I. Hwang, and K. Lee, “Hear your face: Face-based voice conversion with f0 estimation,” in *Proc. Interspeech*, 2024. 2

[34] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, “Conformer: Convolution-augmented transformer for speech recognition,” in *Proc. Interspeech*, 2020. 2

[35] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *NeurIPS*, 2017. 3

[36] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” in *NeurIPS 2021 Workshop*, 2021. 3

[37] T. Afouras, J. S. Chung, and A. Zisserman, “Lrs3-ted: a large-scale dataset for visual speech recognition,” *arXiv:1809.00496*, 2018. 3

[38] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Lip reading sentences in the wild,” in *Proc. CVPR*, 2017. 3

[39] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, “Utmos: Utokyo-sarulab system for voicemos challenge 2022,” in *Proc. Interspeech*, 2022. 3

[40] P. Ma, S. Petridis, and M. Pantic, “End-to-end audio-visual speech recognition with conformers,” in *Proc. ICASSP*, 2021. 3

[41] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, “A lip sync expert is all you need for speech to lip generation in the wild,” in *Proc. ACM MM*, 2020. 3

[42] J. S. Chung and A. Zisserman, “Out of time: automated lip sync in the wild,” in *Proc. ACCV*, 2017. 3

[43] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-shot multi-level face localisation in the wild,” in *Proc. CVPR*, 2020. 3

[44] 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 *Proc. ICCV*, 2017. 3

[45] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, “Scaling rectified flow transformers for high-resolution image synthesis,” in *Proc. ICML*, 2024. 3

[46] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” in *NeurIPS*, 2020. 3

[47] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *Proc. ICLR*, 2021. 4