



LightL2S: Ultra-Low Complexity Lip-to-Speech Synthesis for Multi-Speaker Scenarios

Yifan Liang^{1,2}, Kang Yang³, Fangkun Liu^{1,2}, Andong Li^{1,2}, Xiaodong Li^{1,2}, Chengshi Zheng^{*1,2}

¹Institute of Acoustics Chinese Academy of Science, Beijing, China

²University of Chinese Academy of Sciences, Beijing, China

³Jiangsu Key University Laboratory of Software and Media Technology under Human-Computer Cooperation, Jiangnan University, Wuxi, China

liangyifan@mail.ioa.ac.cn, 6233110046@stu.jiangnan.edu.cn, liufangkun@mail.ioa.ac.cn, liandong@mail.ioa.ac.cn, lxd@mail.ioa.ac.cn, cszheng@mail.ioa.ac.cn

Abstract

Lip-to-speech synthesis in the wild remains challenging due to the limited visual information. While self-supervised models have shown promising results in relatively high-quality lip-to-speech synthesis, their computational demands make them impractical for edge devices. To address this issue, we introduce LightL2S, a novel multi-speaker lip-to-speech system designed to achieve ultra-low complexity for edge deployment. To reduce the computational cost, we adopt a much more efficient architecture MoViNet for visual encoder instead of using the conventional ResNet-18. Furthermore, we introduce Zipformer blocks to efficiently learn prosodic information and quantized self-supervised audio representations from the output features generated by MoViNet. Finally, we employ the differentiable digital signal processing vocoder to synthesize speech. Experimental results demonstrate that LightL2S can generate reasonable speech even with a computational complexity of only 0.8 GMacs.

Index Terms: lip to speech synthesis, computational efficiency, speech synthesis, self-supervised learning, DDSP

1. Introduction

Visual modality plays an important role in human communication. When listening to a speaker, we often rely on facial cues to better understand the content. In extremely noisy environments or when speech is unavailable, visual cues become even more crucial for comprehension. Motivated by these observations, lip-to-speech synthesis aiming to generate speech from lip movements has garnered increasing attention in recent years, especially with the rapid development of deep learning [1–10]. By converting silent visual inputs into intelligible speech, this technology has the potential to assist individuals with speech impairments in regaining their ability to communicate. Additionally, it also plays an important role in silent secret communication scenarios and long-distance surveillance.

Achieving high-quality lip-to-speech synthesis in real-world scenarios remains challenging due to the limited information in the visual modality. The task in unconstrained environments, commonly referred to as ‘lip-to-speech in the wild’, was first introduced in SVTS [1]. This system combines 3D convolutions and ResNet-18 [11] as the visual frontend, employs Conformers [12] for sequence modelling, and finally utilizes a neural vocoder to establish a scalable architecture for lip-to-speech synthesis. Experimental results on the LRS3 [13] dataset demonstrate that SVTS can effectively generate intelligible speech in real-world conditions. The ReVISE [2] system adopts

a similar structure but leverages AV-HuBERT [14], which is a self-supervised audio-visual model consisting of a ResNet-18 based visual encoder and Transformer blocks. Specifically, ReVISE utilizes a pre-trained AV-HuBERT to predict discrete self-supervised learning (SSL) units and a unit-HiFiGAN vocoder for speech synthesis. The integration of AV-HuBERT has led to significant improvements in synthesis performance. Similar frameworks, such as Multi-task [3] and Intelligible-L2S [4] also further extend the framework proposed in SVTS, enhancing the intelligibility of the synthesized speech. Despite these advancements, most of existing methods remain computationally expensive and require extensive training data. Specifically, self-supervised models like AV-HuBERT need high computational costs, which is impractical for deployment on edge devices.

In this paper, we propose LightL2S, an efficient lip-to-speech synthesis system that is optimized for edge devices with ultra-low complexity. Motivated by SVTS and its subsequent works, we introduce several key modifications to enhance efficiency while maintaining performance. Firstly, we adopt MoViNet [15] as the visual frontend, due to it can achieve competitive accuracy with low memory usage when compared with computationally expensive ResNet18-based models. For the generation of pitch and discrete SSL units, we replace the Conformer backbone with Zipformer [16], reducing complexity without sacrificing capacity. Finally, we use a differentiable digital signal processing (DDSP) [17] module to generate speech, following NaturalL2S [10]. The DDSP module provides an acoustic inductive bias for natural speech synthesis at light computational cost. The proposed framework balances synthesis quality and computational efficiency, making it suitable for edge-device deployment.

We conduct experiments on the LRS3 dataset, and the results demonstrate that our model achieves competitive performance compared to previous methods. While the intelligibility of the synthesized speech is worse than some state-of-the-art systems, the generated speech remains reasonably natural, and the overall quality is satisfactory. Notably, LightL2S operates with only 0.8 GMacs and requires no text supervision or large-scale data for training. These results suggest that LightL2S offers a promising solution for resource-limited deployment. To the best of our knowledge, this is the first attempt to achieve lip-to-speech synthesis in the wild with such low complexity.

2. Proposed Approach

2.1. Overview

The proposed LightL2S framework contains three main components: the MoViNet visual frontend, the Zipformer backbone,

* Corresponding author

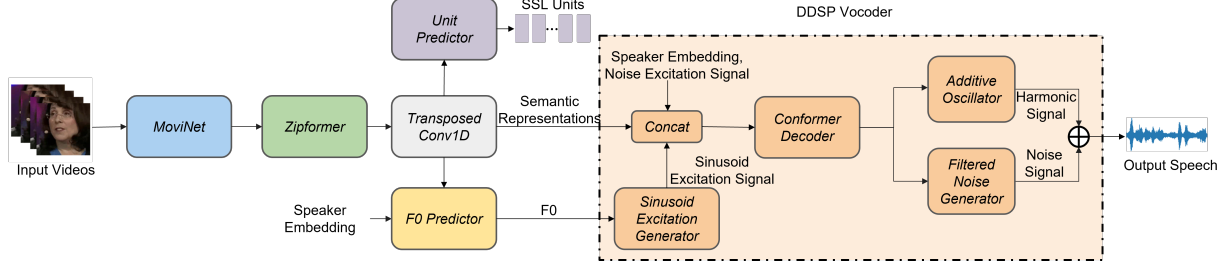


Figure 1: Overview of the proposed LightL2S.

and the DDSP vocoder. The input silent lip moving video is first passed to the visual frontend to extract visual features, and the Zipformer backbone then processes these features, capturing temporal dynamics and learning the mapping from lip movements to self-supervised learning (SSL) units and fundamental frequency (F0) values. These outputs are finally passed to the DDSP vocoder to generate speech. The overall architecture of the LightL2S system is shown in Figure 1. In the following three parts, we describe each component in detail.

2.2. Visual Frontend

Mobile Video Networks (MoViNets) [15] are a family of efficient video recognition frameworks that can reduce memory consumption while maintaining strong performance. Their architecture is optimized by TuNAS [18], which searches for the most efficient structure based on MobileNet-V3 [19]. MoViNets utilize 3D convolutions as the backbone and offer several variants (A0–A6), with A0 having the lowest computational complexity and A6 the highest computational load. To ensure low computational cost while preserving effectiveness, we adopt MoViNet-A0 as the visual frontend and modify it for the lip-to-speech synthesis task. Specifically, we remain the temporal dimension while applying average pooling and adjust the output dimension of the final layer to align with the input size of the Zipformer backbone.

2.3. Zipformer Backbone

Zipformer [16] is a transformer-based model designed to improve computational efficiency, particularly for processing long temporal sequences. It adopts a U-Net-like structure that down-samples frame rates in intermediate blocks, which enables multi-scale temporal modelling. Each Zipformer block is redesigned with two Conformer blocks, where the computed attention weights are reused for computational efficiency. Additionally, Zipformer incorporates BiasNorm to retain length information and introduces two novel activation functions, SwooshR and SwooshL, which accelerate convergence while enhancing performance.

We incorporate prosodic modelling by predicting fundamental frequency (F0) from lip movements, which is because F0 plays a key role in DDSP vocoder-based synthesis by guiding the generation of speech periodicity. The visual features extracted by the visual frontend have a frame rate of 25 Hz, while the SSL units and F0 are sampled at 50 Hz and 100 Hz, respectively. To address this mismatch, we remove the default downsampling operation after the final Zipformer block and introduce a transposed convolution layer to upsample the output by a factor of four, aligning it with the F0 sampling rate. To ensure robust F0 prediction for different speakers while maintain-

ing low complexity, we employ a 1D convolution conditioned on a reference speaker embedding.

Additionally, we generate semantic representations by introducing SSL units as content supervision following [2, 4]. The Unit predictor first downsamples the upsampled features to match the sampling rate of SSL units and then predicts the SSL units through a fully connected layer. These predicted SSL units are supervised with cross-entropy loss against, where the target SSL units are derived through K-means clustering (with 200 clusters) applied to the sixth-layer features of HuBERT-BASE [20]. Previous works [21, 22] have demonstrated that these discrete representations effectively capture phoneme-level semantic information while eliminating speaker-specific characteristics. Moreover, SSL-based supervision removes the need for paired text annotations, making it more practical for low-resource deployment scenarios.

We optimize our model using smoothed cross-entropy loss for the predicted units and L1 loss for the predicted F0 sequences, which can be given by:

$$\mathcal{L}_{\text{unit}} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C q_{i,j} \log \left(\frac{\exp(U_{i,j})}{\sum_{k=1}^C \exp(U_{i,k})} \right), \quad (1)$$

$$q_{i,j} = (1 - \alpha)y_{i,j} + \frac{\alpha}{C}, \quad (2)$$

$$\mathcal{L}_{\text{F0}} = \|\text{F0} - \hat{\text{F0}}\|_1, \quad (3)$$

where N is the number of frames for SSL units, and C is the unit vocabulary size. $U_{i,j}$ represents the predicted logit for the j -th speech unit of the i -th frame, and $y_{i,j}$ denotes the corresponding one-hot label derived from ground-truth speech. The target distribution $q_{i,j}$ is obtained by label smoothing with a smoothing factor α , which is set to 0.1 in our experiments. The F0 and $\hat{\text{F0}}$ represent the ground-truth and predicted values, respectively.

2.4. DDSP Vocoder

The DDSP vocoder combines traditional signal processing theory with deep learning through differentiable components, establishing an interpretable paradigm for speech synthesis. Experiments in [10] have demonstrated that the DDSP vocoder is helpful in the lip-to-speech synthesis task. Additionally, it highlights that end-to-end training of the vocoder effectively improves the quality of synthesized speech. Motivated by these results, we adopt DDSP as the vocoder and train the entire model in an end-to-end manner.

Based on the source-filter model, the DDSP vocoder generates the final speech by decomposing it into periodic and aperiodic parts. In our experiment, the periodic excitation signal

Table 1: *Experimental results on LRS3 dataset. The \uparrow and \downarrow indicate that higher or lower values are better, respectively. The * indicates that the methods with reference speaker embedding.*

Method	UTMOS \uparrow	WER \downarrow	SECS \uparrow	LSE-C \uparrow	GMacs
VCA-GAN [23]	1.31	89.6	0.42	5.23	29.55
Multi-task [3]	1.29	56.8	0.45*	5.20	14.62
DiffV2S [5]	3.07	35.8	0.63	7.21	–
IntelligibleL2S [4]	2.70	27.7	0.76*	7.94	32.34
NaturalL2S [10]	3.66	30.4	0.63	8.08	34.23
LightL2S	2.93	64.8	0.72*	8.01	0.80
Ground Truth	3.59	0.71	–	7.63	–

is generated from the predicted F0 and the aperiodic excitation is sampled by a Gaussian noise. These two excitation signals, along with the speaker embedding and the semantic representations, are input into Conformer blocks to estimate the parameters for speech synthesis: harmonic amplitudes, harmonic phases, noise magnitude, and noise phase.

The synthesis process of the periodic part involves generating a series of harmonics by modulating the base phase with integer multiples and selectively dropping out any harmonics that exceed the Nyquist frequency. The estimated amplitudes are applied to these harmonics, and their sum yields the raw harmonic signal. The estimated harmonic phase is then used to obtain an all-pass filter to refine the harmonic signal in the frequency domain. In parallel, the noise component is refined through a locally time-variant finite impulse response filter using the estimated noise magnitude and phase parameters. Finally, the harmonic and noise components are summed to produce the final synthesized speech. We apply the multi-resolution Short-Time Fourier Transform(STFT) loss to optimize the vocoder,

$$\mathcal{L}_{\text{STFT}} = \sum_L \sum_{\ell} \sum_k \|X_{\ell,k} - \hat{X}_{\ell,k}\|_1, \quad (4)$$

where $X_{\ell,k}$ and $\hat{X}_{\ell,k}$ denote STFT of the ground-truth and predicted speech at frequency bin k of frame ℓ , respectively. In our experiment, the overlap between adjacent frames is 75% and the window size L is set to $\{64, 128, 256, 512, 1024, 2048\}$.

When only applying the L1 loss on spectrograms, the generated results tend to capture overall trends while ignoring local details, resulting in the high-frequency information loss and leading to a blurry and over smoothed spectrogram. To address this problem, we incorporate an adversarial loss to enhance the perceptual quality of the synthesized speech. However, commonly used discriminators in vocoders, such as the multi-scale discriminator (MSD) and multi-period discriminator (MPD) [24], fail to efficiently improve the speech quality in our preliminary experiments. We observe that the DDSP-generated signals are more easily distinguishable by these time-domain discriminators while the generator is hard to eliminate this gap. To mitigate this issue, following [25], we adopt the multi-resolution spectrogram discriminators [26] consists of six sub-discriminators. Each sub-discriminator employs 2D convolutions to distinguish spectrograms at different resolutions. The corresponding loss function is formulated as follows:

$$\mathcal{L}_{adv} = \sum_{i=1}^6 \mathbb{E}_{(X_{j,k}, \hat{X}_{j,k})} \left(D_i(\hat{X}_{j,k}) - 1 \right)^2, \quad (5)$$

$$\mathcal{L}_D = \sum_{i=1}^6 \mathbb{E}_{(X_{j,k}, \hat{X}_{j,k})} \left[\left(D_i(X_{j,k}) - 1 \right)^2 + \left(D_i(\hat{X}_{j,k}) \right)^2 \right], \quad (6)$$

where $D_i(\cdot)$ denotes the output of the i -th sub-discriminator and \mathcal{L}_D is the adversarial loss for discriminator.

Our final loss for training the entire generator of LightL2S is $\mathcal{L}_{\text{total}} = \lambda_{\text{STFT}} \mathcal{L}_{\text{STFT}} + \lambda_{\text{unit}} \mathcal{L}_{\text{unit}} + \lambda_{\text{F0}} \mathcal{L}_{\text{F0}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}}$, where λ_{unit} , λ_{F0} , and λ_{adv} are hyperparameters that balance the contributions of the different loss terms. In our experiments, we set $\lambda_{\text{STFT}} = 45$, $\lambda_{\text{unit}} = 5$, $\lambda_{\text{F0}} = 20$, and $\lambda_{\text{adv}} = 5$.

3. Experiments

3.1. Datasets

We conduct our experiments on the LRS3 dataset, which contains over 400 hours of spoken videos from TED and TEDx videos. It contains around 150,000 utterances with a vocabulary of over 50,000 words and has been one of the most popular corpora in lip reading and lip-to-speech research fields. We follow the splits from the original paper in our experiment.

3.2. Data Preprocessing

We first extract the 68 facial landmarks using dlib [27] and align the face images to a mean face shape. Then, we convert the image into grayscale and normalize each frame. Finally, we crop an 88×88 mouth region for model input. In the audio stream, the audio data is sampled at 16kHz and the corresponding mel-spectrogram is obtained with 80 mel-filterbanks, 640 window size, and 160 hop size. The ground-truth fundamental frequency is obtained by the RAPT algorithm [28] with the window size of 640 and hop size of 160. The reference speaker embedding is extracted through the ESPnet [29], which is a deep learning toolkit for speech processing.

3.3. Model Configuration

As discussed in Section 2.2, we use MoViNet-A0 as the visual frontend and the output dimension is set to 192. For the Zipformer backbone, we follow the Zipformer-S configuration with 6 stacks, each containing 2 blocks. The embedding dimensions for Zipformer blocks are set to $\{192, 256, 256, 256, 256, 256\}$, while the feed-forward dimensions are configured as $\{512, 768, 768, 768, 768, 768\}$. The multi-head attention mechanism uses $\{4, 4, 4, 8, 4, 4\}$ attention heads per block, and the convolutional layers have kernel sizes of $\{31, 31, 15, 15, 15, 31\}$. The DDSP vocoder consists of three Conformer blocks, each with a hidden dimension of 256 and eight attention heads in the multi-head attention module. The DDSP vocoder generates 32 harmonics, with both the noise magnitude and noise phase represented by vectors of dimension 256. For the discriminators, the input multi-scale spectrograms are computed at six different scales, with the number of filters at each scale set to $\{64, 128, 256, 512, 1024, 2048\}$.

3.4. Training Details

We train our model on a single NVIDIA A100 40GB GPU using the AdamW [30] optimizer with $\beta_1 = 0.8$, $\beta_2 = 0.99$ and weight decay of $\lambda = 0.01$. The initial learning rate is set to $5e-4$, with an exponential learning rate decay for adaptive optimization. During training, we randomly sample frames from video sequences, limiting the maximum sequence length to 100 frames and the minimum to 25 frames. To optimize GPU memory usage, waveform segments are restricted to 16,000 samples, with corresponding semantic representations and predicted F0 randomly selected. The model is trained for 100 epochs using BF16 mixed precision. Specifically, we pre-train the model

Table 2: MOS results on LRS3 dataset.

Method	Naturalness \uparrow	Intelligibility \uparrow	Similarity \uparrow
VCA-GAN [23]	1.54 \pm 0.33	1.89 \pm 0.51	1.47 \pm 0.28
Multi-task [3]	1.82 \pm 0.39	2.77 \pm 0.34	1.90 \pm 0.41
DiffV2S [5]	3.69 \pm 0.46	3.49 \pm 0.33	3.25 \pm 0.30
IntelligibleL2S [4]	3.29 \pm 0.43	3.72 \pm 0.31	3.43 \pm 0.27
NaturalL2S [10]	4.10 \pm 0.31	3.98 \pm 0.29	3.51 \pm 0.27
LightL2S	3.63 \pm 0.45	3.20 \pm 0.34	3.72 \pm 0.25
Ground Truth	4.51 \pm 0.24	4.51 \pm 0.22	–

without the discriminator for 80 epochs and subsequently fine-tune it with the discriminator for an additional 20 epochs.

3.5. Evaluation Metrics

Both objective and subjective metrics are chosen to evaluate the performance of the proposed model. For objective evaluation, we assess the quality of the synthesized speech using non-intrusive metrics UTMOS [31]. To measure the intelligibility of synthesized speech, we compute the word error rate (WER) using Auto-AVSR [32]. We also measure LSE-C with a pre-trained SyncNet [33] to evaluate the audio-visual synchronization confidence. Additionally, we measure Speaker Embedding Cosine Similarity (SECS) between the speaker embeddings extracted by Resemblyzer [34] to evaluate the speaker similarity. For subjective evaluation, we conduct a listening test with 10 participants to simultaneously evaluate the naturalness, intelligibility, and speaker similarity of the synthesized speech. Participants are asked to rate 30 randomly selected samples for each aspect on a scale from 1 to 5.

4. Results

4.1. Objective Evaluation

We evaluate our model against baseline on the LRS3 dataset and report the complexity of each method for 1-second video inputs. Notably, Intelligible-L2S [4], DiffV2S [5], and NaturalL2S [10] use AV-HuBERT, while VCA-GAN [23] and Multi-task [3] do not. It is worth noting that ReVISE is not included due to the lack of publicly available code. Similarly, since DiffV2S has not been released, its complexity are not reported. The results, shown in Table 1, highlight that the proposed LightL2S outperforms methods that do not rely on AV-HuBERT in terms of UTMOS. Our model even surpasses Intelligible-L2S, indicating that LightL2S can synthesize speech with a high degree of naturalness even with ultra-low complexity. However, there remains a gap in WER when compared with Multi-task and other AV-HuBERT-based methods. This gap may be due to: (1) Multi-task relying on paired text annotations, which introduces additional text modality during training, and (2) the differences in model scale and training data compared to AV-HuBERT-based methods. We will study these issues in future. In terms of audio-visual synchronization, LightL2S performs better than most methods except NaturalL2S. Additionally, LightL2S achieves superior speaker similarity compared to all methods except Intelligible-L2S. These results indicate that LightL2S is capable of producing speech that is better synchronized with the video, while also retaining more speaker-specific timbre information from the speaker embedding.

4.2. Subjective Evaluation

The subjective evaluation results are presented in Table 2. Among the compared methods, NaturalL2S achieves the high-

Table 3: Results of the ablation study of the proposed model on the LRS3 dataset.

Method	UTMOS \uparrow	WER \downarrow	SECS \uparrow	LSE-C \uparrow	GMacs
Baseline	2.93	64.8	0.72	8.01	0.80
w/o GAN loss	1.58	67.2	0.63	7.69	0.80
320 hop length	2.55	64.8	0.71	7.81	0.65
Conformer Backbone	2.74	66.3	0.70	7.71	1.09

est scores in both naturalness and intelligibility. Notably, the performance of LightL2S is comparable to IntelligibleL2S and DiffV2S while maintaining an exceptionally low computational cost. Interestingly, although NaturalL2S surpasses Ground Truth in objective speech quality metrics (as shown in Table 1), human listeners still perceive Ground Truth as more natural. This discrepancy highlights the gap between objective and subjective evaluations. Additionally, LightL2S achieves the highest speaker similarity MOS, demonstrating its strong ability to preserve speaker identity. The superior intelligibility of IntelligibleL2S, DiffV2S, and NaturalL2S can be attributed to their utilization of AV-HuBERT. Despite LightL2S exhibiting a higher WER than the Multi-task model in objective evaluations, subjective preference leans towards LightL2S. This highlights the importance of human preference in assessing lip-to-speech methods, especially when these methods are ultimately designed for human users. Subjective evaluation results emphasize the significance of perceptual factors such as naturalness and quality in shaping user preference. In summary, the subjective evaluation results further validate the efficiency of LightL2S.

4.3. Ablation Study

We conduct the ablation study to investigate the impact of the key design choices in our model. The results are shown in Table 3. Firstly, we evaluate the effectiveness of the discriminator, the results show that the multi-resolution spectrogram discriminator can improve the entire performance of the model, particularly in terms of speech quality. Next, we evaluate the impact of the hop length in the DDSP vocoder. While aligning the sampling rate of F0 with the SSL units is intuitively beneficial and helps reduce complexity, the results indicate that a 160 hop length performs better in terms of quality metrics. This is because the 160 hop length can better capture the prosodic information to improve speech quality. Finally, we compare the performance of the Conformer with that of Zipformer backbone. By replacing the Conformer with the Zipformer, the overall performance is improved and the computational complexity is even lower. These results indicate that the Zipformer can effectively capture the temporal information from the visual features.

5. Conclusion

In this paper, we propose LightL2S, an ultra-low complexity lip-to-speech synthesis system method. The proposed model integrates the MoViNet visual frontend, Zipformer backbone, and DDSP vocoder to synthesize speech from lip movements. Experimental results on LRS3 dataset demonstrate that our method achieves competitive synthesis quality while maintaining ultra-low computational complexity. However, the intelligibility of the synthesized speech remains a challenge due to the constraints imposed by the lightweight architecture. In future work, we aim to enhance speech intelligibility by incorporating advanced modelling that can be deployed on edge devices.

6. References

- [1] R. Schoburg Carrillo De Mira, A. Haliassos, S. Petridis, B. W. Schuller, and M. Pantic, "Svts: Scalable video-to-speech synthesis," in *Interspeech 2022*. ISCA, 2022, pp. 1836–1840.
- [2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 795–18 805.
- [3] M. Kim, J. Hong, and Y. M. Ro, "Lip-to-speech synthesis in the wild with multi-task learning," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [4] J. Choi, M. Kim, and Y. M. Ro, "Intelligible lip-to-speech synthesis with speech units," in *INTERSPEECH 2023*. ISCA, 2023, pp. 4349–4353.
- [5] J. Choi, J. Hong, and Y. M. Ro, "Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 7812–7821.
- [6] S. B. Hegde, K. R. Prajwal, R. Mukhopadhyay, V. P. Nambodiri, and C. Jawahar, "Lip-to-speech synthesis for arbitrary speakers in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. Association for Computing Machinery, 2022, pp. 6250–6258.
- [7] S. Hegde, R. Mukhopadhyay, C. Jawahar, and V. Nambodiri, "Towards accurate lip-to-speech synthesis in-the-wild," in *Proceedings of the 31st ACM International Conference on Multimedia*, ser. MM '23. Association for Computing Machinery, 2023, pp. 5523–5531.
- [8] J. Choi, J.-H. Kim, J. Li, J. S. Chung, and S. Liu, "V2sflow: Video-to-speech generation with speech decomposition and rectified flow," *arXiv preprint arXiv:2411.19486*, 2024.
- [9] 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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 10 082–10 099.
- [10] Y. Liang, F. Liu, A. Li, X. Li, and C. Zheng, "Natural2s: End-to-end high-quality multispeaker lip-to-speech synthesis with differential digital signal processing," *arXiv preprint arXiv:2502.12002*, 2025.
- [11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [12] 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," *arXiv preprint arXiv:2005.08100*, 2020.
- [13] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: a large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.
- [14] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," *arXiv preprint arXiv:2201.02184*, 2022.
- [15] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "Movinets: Mobile video networks for efficient video recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 020–16 030.
- [16] Z. Yao, L. Guo, X. Yang, W. Kang, F. Kuang, Y. Yang, Z. Jin, L. Lin, and D. Povey, "Zipformer: A faster and better encoder for automatic speech recognition," in *The Twelfth International Conference on Learning Representations*, 2023.
- [17] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts, "Ddsp: Differentiable digital signal processing," in *International Conference on Learning Representations*, 2019.
- [18] G. Bender, H. Liu, B. Chen, G. Chu, S. Cheng, P.-J. Kindermans, and Q. V. Le, "Can weight sharing outperform random architecture search? an investigation with tunas," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 14 323–14 332.
- [19] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.
- [20] 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 Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.
- [21] 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," *arXiv preprint arXiv:2104.00355*, 2021.
- [22] 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.
- [23] M. Kim, J. Hong, and Y. M. Ro, "Lip to speech synthesis with visual context attentional gan," *Advances in Neural Information Processing Systems*, vol. 34, pp. 2758–2770, 2021.
- [24] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 17 022–17 033.
- [25] Y. Liu, B. Yu, D. Lin, P. Wu, C. J. Cho, and G. K. Anumanchipalli, "Fast, high-quality and parameter-efficient articulatory synthesis using differentiable dsp," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 711–718.
- [26] W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," *arXiv preprint arXiv:2106.07889*, 2021.
- [27] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [28] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (rapt)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.
- [29] J.-w. Jung, W. Zhang, J. Shi, Z. Aldeneh, T. Higuchi, B.-J. Theobald, A. H. Abdelaziz, and S. Watanabe, "ESPnet-SPK: full pipeline speaker embedding toolkit with reproducible recipes, self-supervised front-ends, and off-the-shelf models," *Proc. Inter-speech 2024*, 2024.
- [30] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [31] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "Utmos: Utokyo-sarulab system for voicemos challenge 2022," *arXiv preprint arXiv:2204.02152*, 2022.
- [32] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic, "Auto-avsr: Audio-visual speech recognition with automatic labels," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [33] J. S. Chung and A. 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*. Springer, 2017, pp. 251–263.
- [34] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.