DiVISe: Direct Visual-Input Speech Synthesis Preserving Speaker Characteristics And Intelligibility

Anonymous ACL submission

Abstract

Video-to-speech (V2S) synthesis, the task of generating speech directly from silent video input, is inherently more challenging than other speech synthesis tasks due to the need to accurately reconstruct both speech content and speaker characteristics from visual cues alone. Recently, audio-visual pre-training has eliminated the need for additional acoustic hints in V2S, which previous methods often relied on to ensure training convergence. However, even with pre-training, existing methods continue to face challenges in achieving a balance between acoustic intelligibility and the preservation of speaker-specific characteristics. We analyzed this limitation and were motivated to introduce DiVISe (**Di**rect **V**sual-**I**nput **S**peech Synth**e**sis), an end-to-end V2S model that predicts Mel-spectrograms directly from video frames alone. Despite not taking any acoustic hints, DiVISe effectively preserves speaker characteristics in the generated audio, and achieves superior performance on both objective and subjective metrics across the LRS2 and LRS3 datasets. Our results demonstrate that DiVISe not only outperforms existing V2S models in acoustic intelligibility but also scales more effectively with increased data and model parameters[1].

1 Introduction

Video-to-speech (V2S) synthesis aims to generate speech from silent speaker videos. Despite its challenging nature, V2S remains valuable because it conveys richer speaker information beyond language, such as timbre, pitch, tone, and mood, which can be easier to comprehend compared to text transcribed by visual speech recognition. V2S holds immense potential in applications such as restoring speech for individuals with speech impairments or enabling communication in noisy environments (Prajwal et al., 2020a).

Since V2S needs to generate waveforms that accurately reconstruct both speech content and timbre from silent lip movements, it is considered challenging. The lack of prior knowledge about the speech corresponding to the speaker's video clips can confuse the model during training and lead to poor synthesis results (Hsu et al., 2023; Choi et al., 2023a). Most existing V2S approaches (Mira et al., 2022; Hegde et al., 2023; Kim et al., 2023) require an additional speaker embedding derived from extra audio (e.g., random audio clips from the same speaker) to provide speaker characteristics as prior knowledge. However, these models often face challenges in practical scenarios, as speaker identity or voice samples are typically unavailable, especially during inference.

Recent advances in V2S aim to eliminate reliance on speaker embeddings. A major breakthrough is offered by ReVISE (Hsu et al., 2023), which leverages AV-HuBERT (Shi et al., 2022a) as its visual backbone, pre-trained on a combination of audio-visual modalities. By utilizing the prior knowledge encoded in this visual backbone, ReVISE generates high-quality audio without requiring speaker embeddings. To convert the visual backbone's output into waveforms, ReVISE uses a separate vocoder, Unit-HiFiGAN, based on Lee et al. (2022), which takes discrete acoustic units clustered from an audio encoder (Hsu et al., 2021) as input. However, the generated audio lacks speaker-specific characteristics. Similarly, DiffV2S (Choi et al., 2023a) employs pre-training with AV-HuBERT (Shi et al., 2022a) and introduces a speaker embedding-conditioned diffusion model to generate Mel-spectrograms from silent videos. Although speaker embeddings are not required during inference, they are still necessary during a separate training phase before training the diffusion model.

In this work, we propose **Di**rect **V**sual-**I**nput **S**peech Synth**e**sis (DiVISe), an end-to-end, speaker-

[1] Code and weights will be made publicly available after acceptance of this paper.

embedding-free approach that leverages audio-visual pre-training (Shi et al., 2022a) to directly predict Mel-spectrograms from video frames. Thanks to its end-to-end design and simple training process, DiVISe is easy to scale. Our motivation comes from analyzing the Unit-HiFiGAN vocoder used by ReVISE (Hsu et al., 2023), which reveals that reliance on acoustic units adversely affects the speaker similarity of the generated audio. Notably, DiVISe achieves superior objective and subjective evaluation results on the LRS2 and LRS3 datasets. The main contributions of this work are as follows:

- Our analysis of ReVISE (Hsu et al., 2023) revealed that the loss of speaker characteristics is primarily due to its choice of vocoder. The unit-based vocoder struggles to preserve speaker characteristics, whereas vocoders based on Mel-spectrograms perform significantly better in retaining them.
- We introduced DiVISe, an end-to-end V2S approach that predicts Mel-spectrograms. By leveraging audio-visual pre-training and an effective vocoder, DiVISe is not only efficient to train but also excels at preserving speaker characteristics.
- DiVISe excels in generating high-quality audio. Even with visual-only input, it outperforms existing methods that rely on additional audio inputs in both objective and subjective evaluations. Moreover, DiVISe's performance scales efficiently with increased data or model parameters.

2 Related Work

2.1 Speech Synthesis With Vocoders

Generating audio waveforms directly in speech synthesis is challenging because of the complexity of audio signals; therefore, vocoders are employed to convert acoustic features, such as Mel-spectrograms or hidden units, into audio waveforms. Mel-spectrograms are often used as an intermediate audio feature for prediction targets due to their interpretability and computational efficiency in the continuous frequency domain. Yamamoto et al. (2020) proposed Parallel WaveGAN (PWG), which utilizes multi-resolution loss in GAN training. BigVGAN (gil Lee et al., 2023) is trained on a large-scale dataset and successfully improves audio quality. Kong et al. (2020) introduced HiFi-GAN, which employs a combination of multi-period and multi-scale discriminators to differentiate between the generated audio waveform from the Mel-spectrogram and the original waveform during training. Alternatively, non-parametric methods like the Griffin-Lim algorithm (Griffin and Lim, 1984) provide a less accurate estimation of the waveform from the Mel-spectrogram.

Polyak et al. (2021) proposed a new speech re-synthesis model that successfully converts pre-trained HuBERT units back into audible speech. This approach involves adapting HiFi-GAN (Kong et al., 2020) to enable direct waveform generation from SSL (Self-Supervised Learning) units. Later, Lee et al. (2022) discovered that when trained on LJSpeech (Ito and Johnson, 2017), reliance on speaker embeddings and fundamental frequency can be reduced, thereby streamlining the model's input requirements. Following this insight, Hsu et al. (2023) proposed the ReVISE model, which combines pre-trained AV-HuBERT (Shi et al., 2022a) with an upsampled linear projection module that generates SSL unit predictions. Similar to Lee et al. (2022), the SSL units are labeled by a pre-trained HuBERT model and converted back to the original waveform with the help of the Unit HiFi-GAN model (Hsu et al., 2023).

2.2 Video-To-Speech Synthesis

Early works of V2S requires speaker embedding as input to ensure acoustic intelligibility. Prajwal et al. (2020a) developed Lip2Wav and explored training the model on a mixture of speakers with the help of speaker embedding similar to Jia et al. (2019). However, the multi-speaker capability of Lip2Wav heavily depended on these speaker embeddings, without which the model would suffer performance degradation. Inspired by Lip2Wav (Prajwal et al., 2020a), SVTS (Mira et al., 2022) proposed a V2S system built with ResNet18 (He et al., 2016) and Conformer (Gulati et al., 2020), and trained their method on a combination of 1759 hours of audio-visual data comprising LRS3 (Afouras et al., 2018) and VoxCeleb2 (Chung et al., 2018). However, to adapt their approach to a multi-speaker context, SVTS also relied on speaker embeddings from extra audio clips during both training and inference phases. Following SVTS, Kim et al. (2023); Hegde et al. (2023) incorporated additional textual information and speaker embeddings to further enhance V2S performance. Kim et al. (2023) based their V2S model on SVTS and proposed to align the intermediate feature of their V2S model with the

Figure 1: Comparison of the Mel-spectrograms of audios generated by HiFi-GAN (middle) and Unit-HiFiGAN (bottom). To highlight the active speaking regions that overlap the most with the ground truth speech (top), regions with smaller numerical differences from the ground truth are marked in red.

textual transcription of the speech. Hegde et al. (2023) integrated a lip-to-text model in their training pipeline along with a modified text-to-speech (TTS) model that utilizes speaker embeddings to produce well-synchronized audio waveforms. Despite the progress these approaches represent in V2S synthesis, they have not eliminated the reliance on additional inputs, limiting the practical uses of these V2S methods.

3 Preliminaries: Speaker Characteristics Preservation In Vocoders

In Video-to-Speech synthesis, effective speech recovery requires the simultaneous preservation of both content information and speaker characteristics. Lee et al. (2022); Hsu et al. (2023) proposed using acoustic units as vocoder input for speech recovery, successfully encoding phonetic information to reconstruct speech content. However, Hsu et al. (2023) found that these units fail to preserve speaker characteristics and cannot generate sound that matches the speaker's identity.

In this work, we first illustrate the limitations of the unit-based vocoder in speech synthesis through a straightforward visual comparison of Mel-spectrograms, highlighting the differences in generated speech within the frequency domain. In Figure 1, we compare the Mel-spectrograms of speech generated by the unit-based vocoder, Unit-HiFiGAN (Hsu et al., 2023), and the mel-based vocoder, HiFi-GAN (Kong et al., 2020)[2]. The unit-based vocoder shows little resemblance to the original speech in the frequency domain, while the mel-based vocoder aligns closely with the original frequency components, making it more effective at preserving speaker characteristics. This indicates that the speech generated by the mel-based vocoder matches the original speech more closely than that produced by the unit-based vocoder.

From a quantitative perspective, Table 1 compares the performance of the two vocoders. The unit-based vocoder underperforms across all metrics, particularly in preserving speaker characteristics (SECS and EER scores), indicating that it is not well-suited for speech synthesis. Furthermore, intelligibility metrics (PESQ, STOI, ESTOI) and content accuracy (WER) reinforce the conclusion that the mel-based vocoder is a superior choice for speech synthesis. These results confirm that the mel-based vocoder recovers speaker characteristics more accurately than the unit-based vocoder.

Since the V2S task naturally allows for inferring speaker characteristics from visual cues, and given that ReVISE (Hsu et al., 2023) loses speaker identity due to its unit-based vocoder, we are motivated to use Mel-spectrograms as prediction targets in V2S synthesis instead of acoustic units. By employing the mel-based vocoder for speech recovery, we can better capture visual speaker characteristics and reflect speaker variations in the synthesized speech.

[2]For visualization, frequency components with values greater than 0.0 and differences of less than 0.8 from the ground truth are highlighted in red.

Method	Match		Intell.		
	SECS↑	EER↓	PESQ↑	ESTOI↑	WER↓
Unit-HiFiGAN	0.555	40.52	1.17	0.45	14.28
HiFi-GAN	**0.894**	**22.96**	**2.91**	**0.84**	**7.63**

Table 1: Audio synthesis performance of HiFi-GAN and Unit-HiFiGAN on speaker matching and intelligibility.

4 Methodology

The DiVISe framework, shown in Figure 2, comprises two main steps: first, the V2S frontend generates Mel-spectrograms from silent video frames; then, a neural vocoder converts these Mel-spectrograms into audio waveforms. Because the vocoder is trained solely on audio data, the V2S frontend serves as the key component in the overall training pipeline.

4.1 Model Architecture

The V2S frontend comprises a pre-trained AV-HuBERT encoder (Shi et al., 2022a) integrated with a conformer module. In line with the V2S methodology of Choi et al. (2023a), we employ HiFi-GAN (Kong et al., 2020) as our vocoder. To ensure synchronization between the output length of the V2S frontend and the vocoder, we utilize linear reshaping (Shi et al., 2022a), which upsamples the temporal resolution to align with the Mel-spectrogram. A detailed analysis of the model layout can be found in Section B.

The AV-HuBERT encoder comprises a modified ResNet-18 (He et al., 2016) and a transformer encoder (Vaswani et al., 2017). In our configuration, the AV-HuBERT model features a 12-layer transformer with 1024 embedding dimensions and 16 attention heads per layer. Following the AV-HuBERT encoder, we integrate a conformer module (Gulati et al., 2020) to capture temporal dependencies. This integration aligns the encoder's outputs with the target Mel-spectrogram, enabling a smoother transition between stages. The conformer module utilized in our study consists of four conformer blocks, each equipped with a 256-dimensional attention mechanism and four attention heads.

4.2 Method Description

We define the problem of V2S synthesis as follows. The model is given a sequence of video frames $\mathbf{v} = [\mathbf{v}_1, \mathbf{v}2, \ldots, \mathbf{v}m] \in \mathbb{R}^{T\mathbf{v} \times H \times W \times 1}$, where $T\mathbf{v}$ is the temporal dimension, and H, W are the height and width of the frames, respectively. Our objective is to transform the input visual frames into a corresponding sequence of audio waveforms $\mathbf{a} = [\mathbf{a}_1, \mathbf{a}2, \ldots, \mathbf{a}n] \in \mathbb{R}^{T\mathbf{a}}$, where $T\mathbf{a}$ is the length of the audio sequence. Specific configurations for preprocessing and parameters in this section are elaborated in Section A.

4.2.1 V2S frontend

This section outlines the structure, training, and evaluation of the V2S frontend, which converts visual input into Mel-spectrograms.

Visual Backbone We utilize the AV-HuBERT encoder to process video frames $\mathbf{v}$, generating a sequence of embeddings $\mathbf{e_v} \in \mathbb{R}^{T_\mathbf{v} \times C}$, where C represents the feature dimension. This encoder leverages audio-visual pre-training to provide the necessary prior knowledge for our V2S pipeline.

Upsampling To align the temporal resolution of the visual embeddings $\mathbf{e_v}$ with the Mel-spectrogram's sampling rate, we apply linear upsampling, increasing the resolution from $T_\mathbf{v}$ to $kT_\mathbf{v}$. This produces upsampled embeddings $\mathbf{e_m} \in \mathbb{R}^{kT_\mathbf{v} \times C/k}$.

Conformer Module The upsampled embeddings, $\mathbf{e_m}$, are then passed through a conformer module, which captures long-range dependencies and refines the temporal structure. The output of the conformer is subsequently projected through a linear layer to produce the final Mel-spectrogram $\tilde{\mathcal{S}} \in \mathbb{R}^{kT_\mathbf{v} \times F}$, where F represents the number of frequency bins in the Mel-spectrogram.

Training and Evaluation During training, the generated Mel-spectrogram $\tilde{\mathcal{S}}$ is optimized using L1 loss to closely match the ground truth Mel-spectrogram $\mathcal{S} \in \mathbb{R}^{kT_\mathbf{v} \times F}$, ensuring the acoustic accuracy of the generated audio:

$$L_1 = \|\mathcal{S} - \tilde{\mathcal{S}}\|_1. \tag{1}$$

For evaluation, $\tilde{\mathcal{S}}$ is passed to a downstream neural vocoder $\mathcal{V}(\cdot; \phi)$, which converts it back into a speech waveform $\mathbf{w} = \mathcal{V}(\tilde{\mathcal{S}}; \phi), \mathbf{w} \in \mathbb{R}^{T_\mathbf{a}}$. No gradient is propagated back to the V2S frontend during this process. The neural vocoder parameters, denoted as ϕ, are assumed to be pre-trained and fixed.

4.2.2 Vocoder

In this section, we introduce the two-stage training process for the vocoder: pre-training on audio data and fine-tuning on V2S-generated Mel-spectrograms to address distribution mismatch.

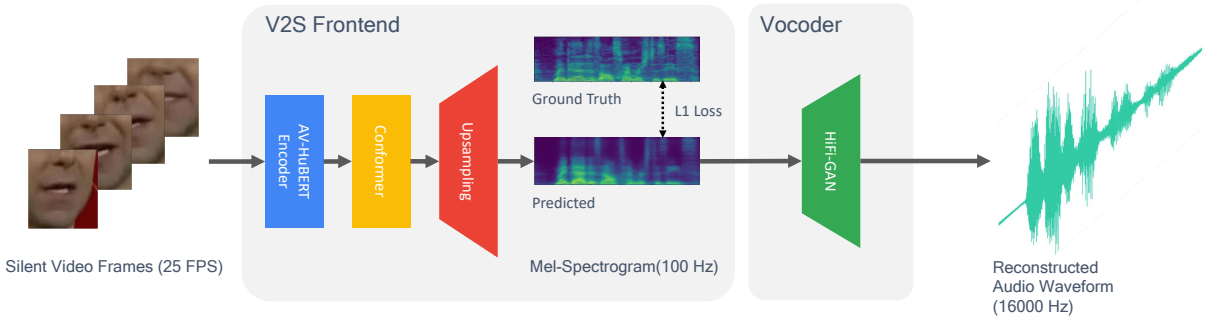

Figure 2: Overview of the DiVISe framework. The V2S frontend maps visual features to Mel-spectrograms, which are then converted into waveforms by the vocoder. During training, DiVISe uses only visual input to generate Mel-spectrograms. During evaluation, the generated Mel-spectrograms are transformed back into audio waveforms.

Pre-training During the pre-training phase, the vocoder $\mathcal{V}(\cdot;\phi)$ is trained on an audio-only dataset consisting of ground-truth waveforms $\mathbf{w}$ and their corresponding Mel-spectrograms $\mathcal{S}$, sampled from the distribution $\mathbb{P}_a$. The vocoder is optimized to reconstruct $\mathbf{w}$ from $\mathcal{S}$ by maximizing the following expectation:

$$\max_{\phi} \mathbb{E}_{\mathcal{S}\sim\mathbb{P}_a}\left[\mathcal{V}(\mathcal{S};\phi) \rightarrow \mathbf{w}\right] \tag{2}$$

Fine-tuning When the V2S frontend generates Mel-spectrograms $\tilde{\mathcal{S}}$, they follow a different data distribution $\mathbb{P}_f$ compared to the vocoder's training distribution $\mathbb{P}_a$, which poses a challenge for accurate waveform recovery. Therefore, we fine-tune the vocoder on the $\tilde{\mathcal{S}}$ generated by the V2S frontend, adjusting its parameters to handle the distribution shift and improve waveform reconstruction. This fine-tuning process is formalized as:

$$\max_{\phi} \mathbb{E}_{\tilde{\mathcal{S}}\sim\mathbb{P}_f}\left[\mathcal{V}(\tilde{\mathcal{S}};\phi) \rightarrow \mathbf{w}\right] \tag{3}$$

Through this adaptation, the vocoder is optimized to account for the distributional differences between real and generated Mel-spectrograms, ensuring consistent and accurate waveform reconstruction from the V2S frontend outputs.

5 Experiments

5.1 Datasets

In our experimental setup, we employ two primary datasets for training and evaluating the V2S models: the Lip Reading Sentences 2 (LRS2-BBC) dataset (Afouras et al., 2022) and the Lip Reading Sentences 3 (LRS3-TED) dataset (Afouras et al., 2018). The LRS3 dataset comprises 433 hours of transcribed English video content, while the LRS2 dataset contains 224 hours. Both datasets feature videos sampled at 25Hz and corresponding audio sampled at 16kHz. For experiments conducted on both datasets, we define the low-resource condition as 30h . The full-resource settings are denoted as 433h for LRS3 and 224h for LRS2, respectively.

Additionally, the LJSpeech Dataset (Ito and Johnson, 2017) is utilized for training the neural vocoders. To train HiFi-GAN and Unit-HiFiGAN, we used a resampled version of the LJSpeech Dataset at 16kHz, following van Niekerk et al. (2021). The VoxCeleb2 dataset (Chung et al., 2018) is also employed but exclusively for evaluating the EER metric using its test set, as speaker pairs for verification trials can be extracted from this dataset.

5.2 Results on Speaker Characteristic Preservation

Speech Similarity and Speaker Verification To test the model's performance in speaker matching, we applied speaker encoder cosine similarity (SECS) (Choi et al., 2023a) and equal error rate (EER) as our measurements, both of which are correlated with speaker embeddings extracted from audio. Higher SECS values and lower EER values indicate better performance in speech similarity and speaker verification, respectively. EER is computed only on VoxCeleb2 (Chung et al., 2018), as in Shi et al. (2022b). A detailed explanation of these metrics can be found in Section D.

Table 4 presents a comparison of SECS scores on the LRS2 and LRS3 datasets. The results for Mira et al. (2022); Kim et al. (2023, 2022); Choi et al. (2023a) are taken from the DiffV2S implementation by Choi et al. (2023a). DiVISe demonstrates leading performance on both datasets, despite not using speaker embeddings during training or evaluation, highlighting its ability to restore speaker characteristics. Additionally, as both DiVISe and

LRS3-TED						
Method	**Data**	**Vocoder**	**PESQ↑**	**STOI↑**	**ESTOI↑**	**WER↓**
Methods taking Visual-only input						
VCA-GAN (Kim et al., 2022)	433h	Griffin-Lim (Kim et al., 2023)	**1.23**	0.474	0.207	96.63
DiffV2S (Choi et al., 2023a)	433h	HiFi-GAN	–	–	0.284	39.2
ReVISE (Hsu et al., 2023)	433h	Unit-HiFiGAN	–	–	–	35.2
Methods requiring extra information						
VAE-GAN †(Hegde et al., 2022)	433h	-	0.51	0.30	0.15	-
AccurateL2S (Hegde et al., 2023)‡	433h	BigVGAN	1.39	0.58	0.37	-
Multi-Task (Kim et al., 2023)‡	433h	Griffin-Lim	1.31	0.497	0.268	66.78
SVTS (Mira et al., 2022)†	433h	PWG	1.25	0.507	0.271	81.5 (Hsu et al., 2023)
SVTS (Mira et al., 2022)†	1759h	PWG	1.26	0.530	0.313	67.4 (Hsu et al., 2023)
Our Implementation						
ReVISE (Hsu et al., 2023)	30h	Unit-HiFiGAN	1.14	0.489	0.288	37.61
ReVISE (Hsu et al., 2023)	433h	Unit-HiFiGAN	1.15	0.491	0.290	36.03
DiVISe	30h	HiFi-GAN	1.21	0.521	0.324	37.29
DiVISe	433h	HiFi-GAN	1.21	**0.529**	**0.333**	**34.52**

LRS2-BBC						
Method	**Data**	**Vocoder**	**PESQ↑**	**STOI↑**	**ESTOI↑**	**WER↓**
Methods taking Visual-only input						
VCA-GAN (Kim et al., 2022)	224h	Griffin-Lim (Kim et al., 2023)	**1.24**	0.407	0.134	109.01
DiffV2S (Choi et al., 2023a)	224h	HiFi-GAN	–	–	0.283	52.7
Methods requiring extra information						
VAE-GAN (Hegde et al., 2022)†	224h	-	0.60	0.34	0.17	-
AccurateL2S (Hegde et al., 2023)‡	224h	BigVGAN	1.47	0.65	0.47	-
Multi-Task (Kim et al., 2023)‡	224h	Griffin-Lim	1.36	0.526	0.341	60.54
IntelligibleL2S (Choi et al., 2023b)†	224h	Hybrid	1.34	0.585	0.412	35.7
SVTS (Mira et al., 2022)†	224h	Griffin-Lim (Kim et al., 2023)	1.34	0.49	0.29	93.91
Our Implementation						
ReVISE (Hsu et al., 2023)	30h	Unit-HiFiGAN	1.12	0.484	0.297	38.60
ReVISE (Hsu et al., 2023)	224h	Unit-HiFiGAN	1.12	0.488	0.302	36.95
DiVISe	30h	HiFi-GAN	1.20	0.529	0.353	39.56
DiVISe	224h	HiFi-GAN	1.19	**0.546**	**0.374**	**34.04**

Table 2: Intelligibility metrics on LRS3 and LRS2 datasets. Top-1/top-2 performances are marked in bold/underlined for methods taking visual input only. Methods marked with †require extra speaker embedding in training, while those marked with ‡require not only audio embedding, but also additional text information. The ASR setting is consistent with Hsu et al. (2023) but may differ in other methods.

DiffV2S (Choi et al., 2023a) utilize the pre-trained AV-HuBERT encoder (Shi et al., 2022a) as the visual backbone, their SECS scores significantly surpass other methods, indicating the effectiveness of audio-visual pre-training in preserving speaker characteristics.

Table 3 shows the comparison on the LRS3 dataset between ReVISE and DiVISe. Compared to ReVISE, DiVISe performs especially better in speaker characteristics preservation, showing advantage in preserving speaker traits in the synthesized audio. This trend aligns with our observations from the vocoder comparison in Table 1, signifying mel-based methods consistently outperform unit-based methods especially in speaker matching.

Method	Match		Intell.		
	SECS↑	**EER↓**	**PESQ↑**	**ESTOI↑**	**WER↓**
ReVISE	0.5384	41.17	1.15	0.290	36.03
DiVISe	**0.6079**	**30.05**	**1.21**	**0.333**	**34.52**

Table 3: Metrics comparison for DiVISe and ReVISE on speaker matching and intelligibility.

Audio-Visual Synchronization LSE (Lip-Sync Expert) (Prajwal et al., 2020b) provides a subjective evaluation of audio-visual synchronization, making it useful for assessing the alignment between synthesized speech and the original video. Higher LSE-C values and lower LSE-D values indicate better synchronization. LSE metrics are calculated following the implementation of Prajwal et al. (2020b); Chung and Zisserman (2016). Table 5 presents the LSE-C and LSE-D evaluation results, which reflect the model's ability in audio-visual synchronization in LRS3 dataset. We compared DiVISe against existing methods on LSE-C and LSE-D scores and have found that DiVISe has achieved superior LSE scores even better than methods requiring speaker embedding as input (Mira et al., 2022; Kim et al., 2023). DiVISe shows a prominent performance in audio-visual synchronization

Method	LRS2	LRS3
w/ audio speaker embedding		
SVTS (Mira et al., 2022)	0.558	0.623
Multi-Task (Kim et al., 2023)	0.525	0.549
w/o audio speaker embedding		
VCA-GAN (Kim et al., 2022)	0.453	0.445
SVTS (Mira et al., 2022)	0.499	0.543
Multi-Task (Kim et al., 2023)	0.457	0.495
DiffV2S (Choi et al., 2023a)	<u>0.581</u>	**0.625**
ReVISE (Hsu et al., 2023)	0.521	0.538
DiVISe	**0.641**	<u>0.608</u>

Table 4: SECS comparison of methods. Top-1/top-2 performances are marked in bold/underlined for methods taking visual input only.

Method	LSE-C↑	LSE-D↓
Methods requiring extra information		
SVTS	6.04†	8.28†
Multi-Task	5.19†	8.89†
Methods requiring visual input only		
DiffV2S	7.28†	7.27†
ReVISE	7.11	7.20
DiVISe (Griffin-Lim)	7.31	6.95
DiVISe (Vocoder)	**7.84**	**6.55**

Table 5: LSE-C and LSE-D results for audio-visual synchronization evaluation. Results reproduced by Choi et al. (2023a) are marked with †.

with either Griffin-Lim or HiFi-GAN vocoder as means of Mel-to-Audio conversion. This elevates lip-sync confidence score suggests that the audio track produced by our method synchronizes well with the speaker lip movements in video, signifying a strong audio-visual correlation.

Human Subjective Speaker Matching Score In this section, we conducted a subjective human evaluation on speaker matching, where participants were required to rate the synthesized speech based on its similarity to the speaker's image. We encourage readers to refer to Section F.2 for more details about this test. The evaluation results are reported in Table 6, with a confidence interval of 95%. The speaker matching scores for mel-based methods significantly outperform those of unit-based methods, indicating greater human satisfaction with the audio-visual connection in speaker characteristics. DiVISe excels in aligning the generated audio with the physical characteristics of the visual speaker's appearance. This framework demonstrates considerable potential in fusing acoustic features from the input video, which are essential for accurately restoring speaker characteristics in the synthesis.

Audio Source	Matching Score
Vocoders	
Unit-HiFiGAN	1.81±0.18
HiFi-GAN	4.64±0.08
V2S Methods	
ReVISE	1.80±0.18
DiVISe	3.90±0.14
Ground Truth	
Source	**4.80±0.07**

Table 6: Subjective measurements of Speech Matching Scores, where the relation of synthesized audio to the physical appearance of speaker in the original video is assessed by test participants.

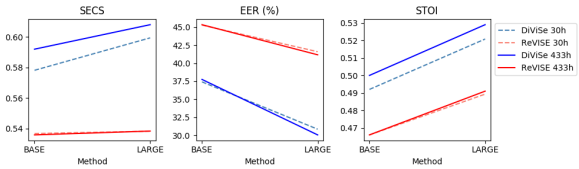

Figure 3: Comparative evaluation results with different model size on LRS3.

5.3 Results on Intelligibility

In this section, we presented the synthesis results for intelligibility in the full resource setting. It is important to highlight that some existing methods incorporate additional information, such as speaker embeddings or text transcriptions, which gives them an advantage over DiVISe, which relies solely on silent video input during both training and evaluation.

Despite this disparity, DiVISe demonstrates remarkable performance improvements, as detailed in Table 2. Specifically, it achieves the lowest WER across both datasets, indicative of its superior capability of speech content recovery outperforming all other approaches. Additionally, DiVISe delivers the highest STOI and ESTOI scores in this setting, indicating superior audio quality and intelligibility. These results underscore DiVISe's robustness, as it performs exceptionally well without requiring additional data. Moreover, DiVISe consistently demonstrates generalizability, particularly on LRS2, which is not part of the AV-HuBERT backbone's pre-training corpus.

5.4 Ablation Studies

In this subsection, we conduct ablation studies on model composition and resource settings.

5.4.1 Analysis on Size of Data and Model

In this section, we analyzed model performance in speaker characteristics preservation and intelligibil-

Pre-train	Method	Match		Intell.	
		SECS↑	EER↓	ESTOI↑	WER↓
✓	DiVISe	**0.6079**	**30.05**	**0.333**	**34.52**
×	DiVISe	0.4842	37.62	0.206	93.98
✓	ReVISE	0.5384	41.17	0.290	36.03
×	ReVISE	0.5304	44.25	0.220	77.24

Table 7: Ablation on audio-visual pre-training.

Method	Match		Intell.	
	SECS↑	EER↓	ESTOI↑	WER↓
DiVISe	**0.6079**	30.05	**0.333**	**34.52**
- Conformer	0.5960	**30.02**	0.322	37.22
ReVISE	0.5384	41.17	0.290	36.03
+ Conformer	0.5393	41.69	0.291	35.67

Table 8: Ablation on Conformer module.

Fine-tune	Match		Intell.	
	SECS↑	EER↓	ESTOI↑	WER↓
✓	**0.6079**	**30.05**	0.333	**34.52**
×	0.5392	32.31	**0.337**	34.99

Table 9: Ablation on vocoder fine-tuning for DiVISe.

ity under varying model size and training resource. The corresponding visualizations of these results are presented in Figure 3.

Size of data DiVISe demonstrates superior performance across all metrics in the full resource setting (433 hours) compared to the low resource setting (30 hours), consistently outperforming ReVISE in speaker matching. As shown in Table 2, despite having a smaller training corpus, DiVISe exceeds most reported methods in intelligibility, even when those methods are trained on larger datasets. Moreover, compared to ReVISE, DiVISe's performance scales up more profoundly with more data, which suggests that DiVISe has significant potential for scaling effectively to larger datasets.

Size of model DiVISe benefits significantly from increased model parameters. In Figure 3, we compare the performance of AV-HuBERT BASE and LARGE (the default setup in our work) as visual backbones, with LARGE being three times larger than BASE. Compared to ReVISE, DiVISe shows more substantial performance improvements across all metrics as the model size increases. This finding suggests that DiVISe achieves greater performance gains as it scales to larger models, making it well-suited for broader V2S applications that require support for a larger number of model parameters.

5.4.2 Effect of Audio-Visual Pre-training

Pre-training is essential for both DiVISe and ReVISE, as shown by the performance gains in Table 7 with pre-trained AV-HuBERT. Pre-training results in significant improvements across all metrics for DiVISe, while ReVISE primarily benefits in intelligibility but fails to meaningfully improve speaker preservation. This disparity underscores ReVISE's limitations in retaining speaker characteristics, which remain unrecovered despite the knowledge acquired from pre-training.

5.4.3 Effect of Conformer Module

As shown in Table 8, integrating the conformer module into DiVISe improves acoustic intelligibility, demonstrating its ability to analyze temporal visual dynamics. In contrast, ReVISE (Hsu et al., 2023) shows more limited improvement with the conformer. For both methods, the integration mainly boosts intelligibility, with little effect on preserving speaker characteristics.

5.4.4 Effect of Vocoder Fine-tuning

Table 9 highlights the effect of fine-tuning the vocoder on V2S frontend outputs. Fine-tuning leads to significant improvements, particularly in speaker matching scores, underscoring its importance for DiVISe. In contrast, intelligibility metrics show minimal impact from fine-tuning, likely due to their robustness to slight frequency variations.

6 Conclusion

In this study, we introduced DiVISe, a V2S model directly predicting Mel-spectrograms from silent videos without relying on any additional inputs both training and inference. By leveraging mel-based vocoder and audio-visual pre-training, DiVISe achieves superior results in speaker characteristics preservation, as confirmed through both subjective and objective evaluations. Besides, DiVISe enhances speech intelligibility, accurately recovering the audible content of speech. Additionally, our findings indicate that DiVISe scales more efficiently with larger datasets and model sizes.

Overall, DiVISe delivers high-quality, intelligible speech with minimal input requirements, highlighting its potential for broader applications in speech synthesis and related fields. In the future, DiVISe can be applied to larger datasets and enhanced with increased model parameters.

7 Limitations

Following Shi et al. (2022a), our current work requires preprocessing the input frames as detailed in Section A.1 of appendix. This limits the application of DiVISe when applied in real-time V2S scenario, especially when locating the area of interest centered around the mouth for each input video frame. This could lead to extra lag in real-time application, and we expect potential streamlining in current preprocessing pipeline to make it more suited towards real-time V2S synthesis in the future.

DiVISe is only tested on LRS2 and LRS3 datasets containing English-only corpus. It is not tested on multi-lingual audio-visual corpus in this work due to the limitation of computational resources, but we still expect to carry out future research on its capability of adapting to other languages (e.g. Chinese, French).

References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2022. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8717–8727.

Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Lrs3-ted: a large-scale dataset for visual speech recognition. *ArXiv*, abs/1809.00496.

Jeongsoo Choi, Joanna Hong, and Yong Man Ro. 2023a. Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding. *Preprint*, arXiv:2308.07787.

Jeongsoo Choi, Minsu Kim, and Yong Man Ro. 2023b. Intelligible lip-to-speech synthesis with speech units. *Preprint*, arXiv:2305.19603.

J. S. Chung and A. Zisserman. 2016. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *Interspeech*.

Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023. Bigvgan: A universal neural vocoder with large-scale training. *Preprint*, arXiv:2206.04658.

D. Griffin and Jae Lim. 1984. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. *ArXiv*, abs/2005.08100.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sindhu B. Hegde, Rudrabha Mukhopadhyay, C. V. Jawahar, and Vinay Namboodiri. 2023. Towards accurate lip-to-speech synthesis in-the-wild. *Proceedings of the 31st ACM International Conference on Multimedia*.

Sindhu B. Hegde, K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2022. Lip-to-speech synthesis for arbitrary speakers in the wild. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22. ACM.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi. 2023. Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech regeneration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18796–18806.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2019. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Preprint*, arXiv:1806.04558.

Minsu Kim, Joanna Hong, and Yong Man Ro. 2022. Lip to speech synthesis with visual context attentional gan. In *Neural Information Processing Systems*.

Minsu Kim, Joanna Hong, and Yong Man Ro. 2023. Lip-to-speech synthesis in the wild with multi-task learning. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Davis E. King. 2009. Dlib-ml: A machine learning toolkit. *J. Mach. Learn. Res.*, 10:1755–1758.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Preprint*, arXiv:2010.05646.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, Juan Pino, and Wei-Ning Hsu. 2022. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339, Dublin, Ireland. Association for Computational Linguistics.

Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2022. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 4(11):930–939.

Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W. Schuller, and Maja Pantic. 2022. Svts: Scalable video-to-speech synthesis. *Preprint*, arXiv:2205.02058.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdel rahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations. In *Interspeech*.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay Namboodiri, and C V Jawahar. 2020a. Learning individual speaking styles for accurate lip to speech synthesis. *Preprint*, arXiv:2005.08209.

K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. 2020b. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, New York, NY, USA. Association for Computing Machinery.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdel rahman Mohamed. 2022a. Learning audio-visual speech representation by masked multimodal cluster prediction. *ArXiv*, abs/2201.02184.

Bowen Shi, Abdel rahman Mohamed, and Wei-Ning Hsu. 2022b. Learning lip-based audio-visual speaker embeddings with av-hubert. *ArXiv*, abs/2205.07180.

Benjamin van Niekerk, Marc-André Carbonneau, Julian Zaïdi, Matthew Baas, Hugo Seuté, and Herman Kamper. 2021. A comparison of discrete and soft speech units for improved voice conversion. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6562–6566.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Preprint*, arXiv:1706.03762.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2020. Self-training and pre-training are complementary for speech recognition. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203.

A Training Details

A.1 Preprocessing Details

DiVISe has adopted the data processing pipeline from AV-HuBERT (Shi et al., 2022a) for data consistency as the upstream model is a pre-trained AV-HuBERT model. The area of interest, specifically a 96×96 pixel region centered around the mouth, is then extracted from each frame[3]. In training, we further process the images by extracting a smaller 88×88 region from the larger mouth-centered area. The selected frames are subject to a 50% chance of being horizontally flipped to enhance the robustness of our model. In training we limit the length of each sampled video clip to 4.0 seconds starting from the beginning of the video. At inference, the entire video is always loaded into the model, except for EER evaluation. We consistently perform a center crop of the 88×88 region without applying any horizontal flipping in evaluation.

A.2 Parameters Setting

For LRS2 and LRS3 dataset, h_0 is set to 25 to match the sampling frequency and the final waveform is at a frequency of 16kHz. We follow the established practice of van Niekerk et al. (2021) by extracting the 128-dimensional Mel-Spectrogram from the raw audio waveform at 10-ms intervals for both the LRS3 and LJSpeech datasets.

We choose a hop size that aligns $\mathcal{S}$ with $\tilde{\mathcal{S}}$ at the same sampling frequency (kh_0). Hop size is set to be 160 to produce Mel-spectrogram at 100Hz. In this way, the upsampled visual features can match with the sampling frequency of the Mel-spectrogram ($k \times 25 = 100$Hz).

A.3 Training And Evaluation Sets

We follow the setup of AV-HuBERT (Shi et al., 2022a) to determine different sets of datasets in hours. The LRS2 is partitioned into three sections (pre-train for 195 hours, trainval for approximately 30 hours, and test for 0.5 hours), with a predefined

[3]We applied the dlib tool (King, 2009) to each video clip in order to detect the 68 facial keypoints. Subsequently, we employ an affine transformation to align each video frame with a standard reference face frame.

validation set. The LRS3 dataset is segmented into three splits (pre-train for around 400 hours, trainval for 30 hours, and test for 1 hour), with the validation set being extracted from the same subset of the LRS3 trainval split as identified by Shi et al. (2022a).

The 30h setting employs only the trainval split (excluding the validation set) from the LRS2 and LRS3 dataset for model training, while the 433h and 224h setting encompasses the combined pre-train and trainval splits for LRS3 and LRS2, respectively. Regardless of the resource setting applied to each dataset, the validation and test sets remain consistent during the model evaluation for each dataset, with the sole exception of EER evaluation, which only uses test pairs of trials for each item in the VoxCeleb2 test set.

A.4 Experimental Configurations

A.4.1 V2S frontend

In our experimental setup, DiVISe leverages the AV-HuBERT LARGE model (Shi et al., 2022a), which is pre-trained for 5 epochs on the LRS3 dataset and the English-only subset of VoxCeleb2 (Chung et al., 2018), comprising a total of 325 million parameters[4]. This choice aligns with the setting of Hsu et al. (2023) and serves as the default configuration for the experiment. For comparison, the AV-HuBERT BASE model (Shi et al., 2022a) with 103 million parameters, is pre-trained solely on LRS3. This yields a relatively modest enhancement for our method. The conformer module is modified from the implementation of Ma et al. (2022), with the visual frontend removed.

For both ReVISE and DiVISe, V2S frontend training was accomplished on 8 GPUs, with each GPU requiring fewer than 45K updates for about 26 hours. Models are selected based on the lowest L1 loss of Mel-spectrogram and highest accuracy of unit prediction for DiVISe and ReVISE respectively. Specific configurations on LRS2 and LRS3 dataset can be seen in Table 10 and Table 11. Table 10 exhibits full resource setting and low resource setting, with their differences specified in Table 11. The total number of updates may be slightly less than those detailed in Table 11, due to the model selection criteria that we applied. We follow the experimental settings of Hsu et al. (2023) for the full resource setting, but reduce the number of total updates to 1/8 and the number of GPUs to 1/2 to avoid overfitting the volume of data in the low resource setting. The tri-stage LR schedule refers to the learning rate schedule that increases linearly from 0 in the first stage, constantly staying at the maximum learning rate value in the second stage, and finally linearly decreasing to 5% of the peak learning rate value in the last stage. The LR schedule evolves on the per-update level. We freeze the pre-trained AV-HuBERT module for a few updates at the beginning of training. The batch size / GPU means the number of video samples loaded to each GPU and is effectively equivalent in size to a batch size of 1000 video frames per GPU reported in Hsu et al. (2023) due to a max sample length of 4s with 25 FPS of video clips.

tri-stage LR schedule	(10%, 20%, 70%)
peak learning rate	6e-5
batch size / GPU	10
max sample length	4s
Adam (β_1, β_2)	(0.9, 0.98)

Table 10: Experimental settings for LRS2 and LRS3.

Settings of Resource	**Full**	**Low**
num. of updates / GPU	45000	11250
num. of frozen steps	5000	1250
num. of GPU	8	4

Table 11: Experimental settings for different resources.

A.4.2 Vocoder

For DiVISe, we trained HiFi-GAN with 128 Mel-frequency bins as the neural vocoder to ensure consistency with the ReVISE framework (Hsu et al., 2023). HiFi-GAN is first trained on the resampled LJSpeech dataset for 400k updates with a single GPU. We then fine-tune HiFi-GAN on Mel-spectrograms generated by upstream V2S frontend trained on LRS3 and LRS2 datasets, respectively, for 20k updates (per GPU) with 8 GPUs. The training code is developed based on the implementation of van Niekerk et al. (2021).

For ReVISE implementation, we align with the settings of Hsu et al. (2023). We trained a 2000-class K-Means model from feature extracted in the last layer of the HuBERT BASE model pre-trained on LibriSpeech 960h dataset in the third

[4]The pre-trained weights used in this work are publicly accessible at https://facebookresearch.github.io/av_hubert/

Modules	DiVISe	DiVISe-BASE	ReVISE
Visual Backbone	325M	103M	325M
Upsampling	-	-	8M
Conformer	11M	11M	-
Sum w/o Vocoder	336M	114M	333M
Vocoder	14M	14M	15M
Sum w/ Vocoder	350M	128M	348M

Table 12: Parameters comparison of DiVISe in model size. DiVISe-BASE replaces the visual backbone with pre-trained AV-HuBERT BASE model, trimming model size to 1/3.

epoch[5]. We then train Unit-HiFiGAN on units labeled by this K-Means model for 400k updates (per GPU) on the resampled LJSpeech dataset with 8 GPUs. Since the unit vocabulary is shared with its upstream V2S frontend in prediction, there is no need for further fine-tuning.

B Model Sizes

From Table 12, one can see that the inclusion of the conformer module in our proposed model adds merely 10.29 million parameters, which is relatively minimal when compared to the overall size of the model. This demonstrates that the conformer module does not significantly increase the model's size. DiVISe worked around the upsampling module by directly applying linear reshaping following Shi et al. (2022a). Substituting the visual backbone with AV-HuBERT BASE results in an even more compact version of the proposed model (DiVISe-BASE), suitable for scenarios requiring a lighter computational footprint. In comparision, DiVISe maintains a relatively moderate size.

C Subjective Mean Opinion Score Evaluation

Audio Source	MOS Score
Unit-HiFiGAN	4.37 ± 0.11
HiFi-GAN	4.24 ± 0.12
Ground Truth	4.47 ± 0.11

Table 13: Subjective MOS score measurements for vocoders on LRS3 dataset.

We conducted a subjective evaluation for Unit-HiFiGAN and HiFi-GAN to further explore subjective human feedback on their synthesized audio clips. The confidence interval for Mean opinion score (MOS) scores is 95%. Other settings can be found in Section F.2. In Table 13, there is a slight drop in the MOS score for HiFi-GAN compared to Unit-HiFiGAN. We assume that this performance gap is attributed to the relatively more consistent and neutral style of audio clips in all samples generated by unit-based methods, without considering the identity of the speaker. Our experiment reveals that this trait of unit vocoder gains a better impression from human listeners. However, it is important to note that no speaker-related information is given to test participants to assess how related synthesized audio clips are to the visual speaker, which is an in-born drawback for this audio-only MOS evaluation.

D More Details on Metrics

In this section, we provide detailed explanation and implementation for metrics mentioned in this work for reader's better understanding.

Speech Similarity and Speaker Verification For speech similarity, SECS is computed by simply averaging the cosine similarity between speaker embeddings of ground-truth and synthesized audio (Choi et al., 2023a). On the other hand, EER is a metric used to assess the performance of speaker verification. For EER evaluation, we first set up a list of positive and negative pairs for every sample in the VoxCeleb2 test set that sums up to a total of 72,474 pairs of trials, similar to Shi et al. (2022b). To compute EER, we first synthesize audio paris with V2S methods given any video pair of trials, then we use the cosine similarity of speaker embeddings computed from the generated audio clips for every pair of trials as the matching score. The matching score serves as the prediction score, where ground-truth labels are assigned as 1 for positive pairs and 0 for negative pairs. These scores are used to compute false acceptance rates and false rejection rates. To simplify EER calculation, we use only one pair of 4-second audio clips for each pair of trials. For both SECS and EER, we use an existing pre-trained audio speaker encoder (Jia et al., 2019) to gain speaker embedding in a setup identical to Choi et al. (2023a).

Intelligibility The intelligibility of synthesized audio clips is quantitatively assessed by objective metrics, specifically Short-Time Objective Intelligibility (STOI), Extended STOI (ESTOI) and Percep-

[5]The pre-trained HuBERT model used for replicating ReVISE in this study is obtained from https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt

tual Evaluation of Speech Quality (PESQ)[6]. These metrics provide a comprehensive understanding of the speech intelligibility and quality that our model is capable of producing.

Additionally, the synthesized waveforms are further transcribed into text using the publicly available wav2vec 2.0 model pre-trained by Xu et al. (2020)[7] identical to the choice of ReVISE (Hsu et al., 2023). This ASR system allows us to evaluate the intelligibility of the generated speech in terms of its WER (the lower the better) with the text transcribed from ground-truth audio. The transcription results also serve as a proxy for how well the model's output would be understood in real-world applications, where accurate conversion of speech to text is often crucial. This ASR reports a WER of 5. 68% in the LRS3 dataset in our setting.

E Performance on various text lengths

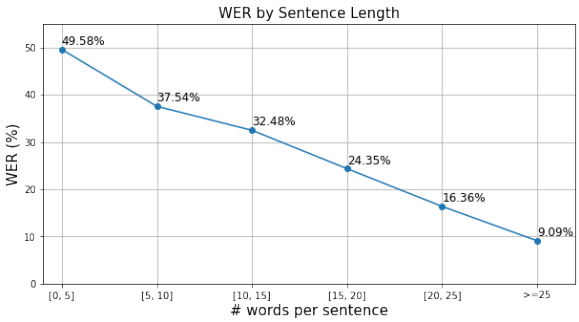

Figure 4: WER reported on different length spans of Ground-Truth texts. Video clips corresponding to longer lengths generally have better WER performance for DiVISe.

The graph in Figure 4 presents the WER by sentence length of the text transcribed from ground-truth audio for DiVISe. DiVISe exhibits a decrease in WER as the length of the sentence increases. A similar observation is also found by Shi et al. (2022a), where WER decreases as the length of ground-truth text increases in lip-reading (VSR) task. DiVISe gains better performance at handling longer sequences of words evaluated by ASR in the V2S setting, possibly due to the availability of more contextual information within the video that aids in speech synthesis. This could mean that, intuitively, V2S and VSR are two highly correlated tasks that share similar traits.

F Research Ethics and Volunteer Testing

F.1 Use of Artifacts

For all datasets used in this work (LRS3-TED (Afouras et al., 2018), LRS2-BBC (Afouras et al., 2022), VoxCeleb2 (Chung et al., 2018), and LJSpeech (Ito and Johnson, 2017)), we strictly adhere to the usage guidelines outlined by their respective authors. The datasets we use are publicly available online and do not contain any personally identifying information or offensive content. No additional steps were required for anonymization or content screening, as the data has already been curated and verified for research purposes by the original providers. No other data has been collected, repackaged, or released as part of this work.

For the code artifacts used in this work, all relevant code has been appropriately credited through proper citations. Metrics, results, and other contributions from external authors have been thoroughly acknowledged and referenced. The use of existing artifacts in this work aligns with their intended purpose as specified by their original creators. For the artifacts we generate, their intended use is explicitly for research purposes.

F.2 Human Subjects Setup in Subjective Evaluation

In this study, we recruited 15 student participants to conduct subjective evaluations of the synthesized speech data. Participation was voluntary, with no financial compensation or course credit provided. Each participant received detailed instructions, including a step-by-step guide and screenshots of the evaluation interface, ensuring clarity in the evaluation process. Since the study posed no potential risks to the participants, no specific disclaimers were necessary. Informed consent was obtained prior to participation, with students agreeing to the use of their anonymized evaluation data solely for research purposes. Additionally, participants were explicitly informed that no personal identifying information (PII) would be collected or stored during the study.

In Section 5.2, we performed a subjective speaker matching evaluation, in which each test participant is provided with a head-cropped picture of the speaker in the original video and the

[6]The PESQ metric was calculated using version 0.0.4 of the pesq Python package, whereas STOI and ESTOI metrics were determined with version 0.3.3 of the pystoi Python package.

[7]The wav2vec 2.0 model with its weights used for ASR evaluation is sourced from https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

corresponding ground-truth speech as reference. Based on the given reference, the participants are asked to give their opinion scores on a list of candidate speeches synthesized from difference sources. This subjective evaluation was carried out on 20 randomly selected samples from the LRS3 test set. The full text of instruction given to the test participant is:

> Thank you for your time to take this test! In the following tests, you will be given a speaker image paired with different speech clips. You are only required to score (from 1 to 5) how much you think the given speech matches the speaker image. Speech content and quality do not matter in this test.
>
> We recommend you to put on your earphone for better intelligibility of audio.

In Section C, we also performed subjective MOS test on vocoders on the same range of samples. The test settings align with the setup of Section 5.2. As vocoders should be evaluated in an audio-only setup, no reference speaker image is provided. The full text of instruction is:

> Thank you for your time to take this test! In the following tests, you will be given different speech clips. You are only required to score (from 1 to 5) on the quality of the speech. Slight differences in speech content compared to the reference are normal and can be disregarded.
>
> We recommend you to put on your earphone for better intelligibility of audio.

G Transcription Examples

This section evaluates the intelligibility of the audio generated by our method through a comparative analysis with ground-truth audio transcribed by the ASR system. Table 14 features examples of transcriptions from ground-truth audio and our generated audio, as interpreted by the ASR system.

The examples showcase the effectiveness of DiVISe in accurately replicating ground-truth transcriptions through its audio output. For example, "*were on the green*" closely resembles the ASR transcription "*were on the grid*" from ground-truth audio, demonstrating the high intelligibility of the generated audio. However, there are also notable discrepancies due to unclear articulation in the generated audio, such as "*I think we need*" being interpreted as "*Ethic would lead*". These examples highlight the strengths of DiVISe in generating intelligible audio that can be effectively transcribed by an ASR system.

GT	We're on the grid
ASR (GT)	Were on the grid
ASR (DiVISe)	Were on the green
GT	I was not going to cry myself to sleep
ASR (GT)	I was not going to cry myself to sleep
ASR (DiVISe)	I was not in to crow myself asleep
GT	One woman no longer believes love will ever find her
ASR (GT)	One woman no longer believes love will ever find her
ASR (DiVISe)	Where woman no longer believes lovel ever factor
GT	I think we need something like a Manhattan Project on the topic of artificial intelligence
ASR (GT)	I think we need something like a Manhattan Project on the topic of artificial intelligence
ASR (DiVISe)	Ethic would lead something like a Manhattan Project on the topic of artificial case
GT	And he was talking about the importance of coaching boys into men and changing the culture of the locker room and giving
ASR (GT)	And he was talking about the importance of coaching boys into men and changing the culture of the locker room and giving
ASR (DiVISe)	And he was talking about the importance of coaching poisons a mine and changing the culture of the lakarone and even
GT	So people hear about this study and they're like great if I want to get better at my job I just need to upgrade my browser
ASR (GT)	So people hear about this study and their leg greate if I want to get better at my job I just need to upgrade my browser
ASR (DiVISe)	So people hear about the study and they're like radophoni to get better my job I just said to operate my brother

Table 14: DiVISe has the ability to create audio that yields audio-to-text conversion outcomes similar to ground truth when transcribed by ASR. Differences between transcribed text and ground-truth text are marked in red.