Let There Be Sound: Reconstructing High Quality Speech from Silent Videos

Ji-Hoon Kim[*], **Jaehun Kim**[*], **Joon Son Chung**

Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{jh.kim, kjaehun, joonson}@kaist.ac.kr

Abstract

The goal of this work is to reconstruct high quality speech from lip motions alone, a task also known as lip-to-speech. A key challenge of lip-to-speech systems is the one-to-many mapping caused by (1) the existence of homophenes and (2) multiple speech variations, resulting in a mispronounced and over-smoothed speech. In this paper, we propose a novel lip-to-speech system that significantly improves the generation quality by alleviating the one-to-many mapping problem from multiple perspectives. Specifically, we incorporate (1) self-supervised speech representations to disambiguate homophenes, and (2) acoustic variance information to model diverse speech styles. Additionally, to better solve the aforementioned problem, we employ a flow based post-net which captures and refines the details of the generated speech. We perform extensive experiments on two datasets, and demonstrate that our method achieves the generation quality close to that of real human utterance, outperforming existing methods in terms of speech naturalness and intelligibility by a large margin. Synthesised samples are available at our demo page: https://mm.kaist.ac.kr/projects/LTBS.

Introduction

Have you ever wondered what Charlie Chaplin's movies would sound like if they weren't silent? Indeed, there have been many discussions about what is said in archival silent movies (Smith 1987; Midgley 2006). The ability to reconstruct speech from silent videos opens up interesting applications, such as redubbing silent movies, simulating natural utterance for those who suffer from aphonia, and understanding conversations at a distance.

As a result, the research in lip-to-speech has attracted an increasing amount of attention in recent years (Kumar et al. 2019; Mira et al. 2022; Kim, Hong, and Ro 2023). This line of research has also benefited from the advances in deep learning methods, and in particular, self-supervised learning, since the training leverages natural occurrence of audio and video as a mode of supervision.

A lip-to-speech (LTS) system aims to learn a mapping from silent lip movements to the corresponding speech. This is a challenging one-to-many mapping caused by the

two major obstacles. One is the existence of homophenes, words that have almost identical lip movements but distinct phonemes (e.g. 'bit' and 'pit'). The ambiguity of homophenes brings the one-to-many relationship between lip motions and phonemes (Chung et al. 2017; Kim, Yeo, and Ro 2022). Another obstacle is the multiple variations in speech; same phonemes can be mapped to diverse speech styles based on individual characteristics such as timbre, intonation, and accents (Elias et al. 2021; Kim et al. 2022).

Numerous attempts have been made to improve the quality of LTS systems. Existing works (Le Cornu and Milner 2015, 2017) utilise hand-crafted visual features. However, deep-learning based approaches employ end-to-end methods that generate auditory features directly from silent lip motions. Early deep-learning based methods (Ephrat and Peleg 2017; Kumar et al. 2019) estimate linear predictive coding (LPC) features within a short video clip. Recently, many works (Prajwal et al. 2020; Kim, Hong, and Ro 2021; Yadav et al. 2021; He et al. 2022; Mira et al. 2022; Kim, Hong, and Ro 2023) adopt mel-spectrogram as a regression target because it contains more sufficient acoustic information than LPC. Despite the advances, the previous methods do not fully address the one-to-many mapping issue, suffering from a mispronounced and over-smoothed synthetic speech.

In this paper, we propose a novel LTS system that highly improves the synthetic quality by alleviating the intrinsic one-to-many mapping problem. To disambiguate homophenes, we employ self-supervised speech representations as a condition for linguistic information. Previous studies (Baevski et al. 2020; Hsu et al. 2021) have proved that self-supervised learning (SSL) speech models can acquire rich speech representations without manually labeled text. In particular, it has been demonstrated that the representations from specific layers of the SSL model contain elaborate linguistic information independent of paralinguistic features (wen Yang et al. 2021; Chang, Yang, and Lee 2022). Motivated by this, we explore the intermediate layers of SSL model and utilise the hidden representations to produce accurate content without using text labels.

Moreover, we adopt acoustic variance information such as pitch and energy in order to model diverse speech variations. With the help of the acoustic variations, the model can not only ease the one-to-many mapping but also learn prosody of speech which is a key factor for realistic speech

---

[*]These authors contributed equally.

synthesis (Skerry-Ryan et al. 2018; Sun et al. 2020). To further address the one-to-many mapping, we use a flow based post-net (Ren, Liu, and Zhao 2021) which refines acoustic representations with enhanced modelling capability of capturing fine-grained details (Ren et al. 2022). Combined with the variance information, the post-net helps to learn the complex one-to-many-mapping between phonemes and speech, thereby improving the naturalness of the synthesised speech.

We conduct extensive experiments on two public datasets with qualitative and quantitative evaluation metrics. The results show that the proposed method achieves exceptional generation quality, making a mean opinion score (MOS) gap of only 0.28 in naturalness and 0.16 in intelligibility compared to the *vocoded* speech[1].

In summary, we directly tackle the intrinsic one-to-many mapping problem of LTS, arising from the existence of homophenes and multiple speech variations. We adopt self-supervised speech representations as a linguistic condition to disambiguate homophenes without using text labels. To model speech variations, we employ acoustic variance information which helps to capture diverse speech styles. Experimental results prove that our method outperforms existing approaches, achieving human-like generation quality.

Related Works

Lip-to-Speech

With the rapid development of deep-learning, LTS pipelines have been simplified to end-to-end systems, eliminating the need for hand-crafted visual features (Le Cornu and Milner 2015, 2017). Vid2Speech (Ephrat and Peleg 2017) and Lipper (Kumar et al. 2019) adopt convolutions to estimate low-dimensional LPC features. However, since the LPC features contain insufficient information for producing audible waveforms, recent works mostly utilise mel-spectrogram as a regression target. Taking mel-spectrogram, (Prajwal et al. 2020) modifies sequence-to-sequence TTS model (Shen et al. 2018) to construct an LTS system in a single speaker scenario. VCA-GAN (Kim, Hong, and Ro 2021) and (Yadav et al. 2021) further enhance the quality by taking speech variations into account.

To leverage contextual information, GlowLTS (He et al. 2022) proposes a non-autoregressive network architecture based on transformer (Vaswani et al. 2017), and SVTS (Mira et al. 2022) adopts conformer (Gulati et al. 2020). Most recently, MT (Kim, Hong, and Ro 2023) employs content supervision by using additional text labels to produce accurate content. In contrast to previous studies, we focus on alleviating the one-to-many mapping problem in LTS systems by clarifying homophenes and explicitly modelling speech variations, without using text labels.

Text-to-Speech

Text-to-speech (TTS) systems transform text inputs to intermediate acoustic representations (e.g. mel-spectrogram),

which are then converted into audible waveform by a vocoder (Prenger, Valle, and Catanzaro 2019; Yamamoto, Song, and Kim 2020; Kong, Kim, and Bae 2020; Kim et al. 2021). The TTS systems have evolved over time, shifting from conventional concatenative (Hunt and Black 1996) and statistical parametric approaches (Black, Zen, and Tokuda 2007) to end-to-end systems (Shen et al. 2018; Ren et al. 2021; Łańcucki 2021; Lee et al. 2021b; Popov et al. 2021). While recent TTS systems achieve human-like synthetic quality, they require a large amount of manually annotated transcriptions. LTS systems have attracted increasing attention since they can be trained without transcriptions, but the generation quality largely lags behind that of TTS. In this work, we propose a high-quality LTS method that benefits from the self-supervised nature, while generating natural speech whose quality is comparable to that of TTS.

Self-Supervised Learning

Over the recent years, SSL has been an emerging approach for acquiring comprehensive data representations from unlabeled data, and it has achieved notable success in natural language processing (Kenton and Toutanova 2019) and computer vision (Chen et al. 2020; Lin et al. 2021). In speech processing, wav2vec2.0 (Baevski et al. 2020) and HuBERT (Hsu et al. 2021) demonstrate promising results, providing various applications such as speech recognition (Baevski et al. 2021), voice conversion (Lee et al. 2021a; Choi et al. 2021), and speech resynthesis (Polyak et al. 2021). Following works (wen Yang et al. 2021; Chang, Yang, and Lee 2022) report that the hidden representations from different layers pertain to distinct speech attributes such as voice characteristics and linguistic content. Inspired by this, we leverage self-supervised hidden representations from the specific layer of the SSL model to provide linguistic condition, mitigating homophene problems.

Method

Given a silent talking face video, our goal is to synthesise the corresponding mel-spectrogram. As shown in Figure 1(a), the proposed model mainly consists of three components: video encoder, variance decoder, and flow based post-net. The video encoder extracts distinct visual features from input videos, and the variance decoder successively produces coarse mel-spectrogram conditioned on linguistic features and acoustic speech variations (i.e. pitch and energy). The flow based post-net elaborates the coarse mel-spectrogram into the fine-grained one, and the result is finally converted to an audible waveform by a pre-trained neural vocoder (Kim et al. 2021).

Video Encoder

From the input video with T_v frames, the video encoder extracts distinct hidden representations $\boldsymbol{h}_v \in \mathbb{R}^{T_v \times d}$ where d denotes the hidden embedding dimension. As depicted in Figure 1(b), the video encoder comprises 3D convolution (Ji et al. 2012), ResNet18 (He et al. 2016), and a conformer encoder (Gulati et al. 2020). Combined with 3D convolution,
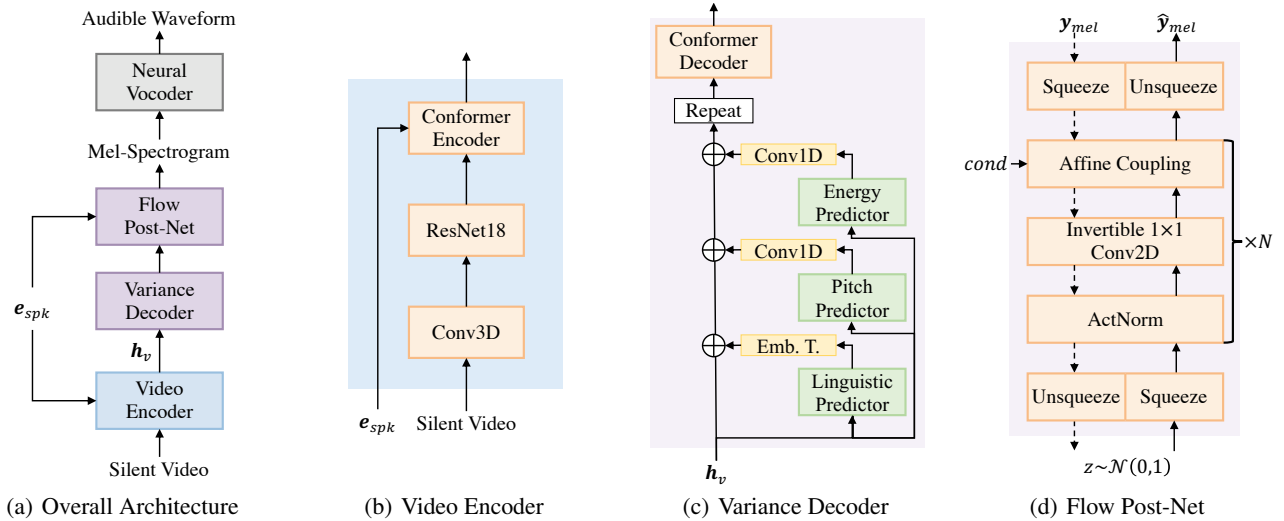
---

[1]*Vocoded* speech refers to speech reconstructed from the ground truth mel-spectrogram through a vocoder, and thus it is practically considered the upper bound quality for our evaluation.

Figure 1: In subfigure (a) and (b), $\boldsymbol{e}_{spk}$ is a speaker embedding. In subfigure (a) and (c), $\boldsymbol{h}_v$ denotes the encoded video feature. In (c), Emb.T. refers to an embedding table. In subfigure (d), the paths with dotted lines are operated only in a training stage. $\boldsymbol{y}_{mel}$ and $\hat{\boldsymbol{y}}_{mel}$ refer to the ground truth and predicted mel-spectrogram, respectively. $cond$ means the post-net conditions which contain the input and output of the conformer decoder, and $\boldsymbol{e}_{spk}$. In our experiment, we set $N = 8$.

ResNet18 can capture adjacent contexts and local visual features, as proved in (Petridis et al. 2018; Ma, Petridis, and Pantic 2022). On the top of ResNet18, we add a conformer encoder which is a stack of conformer layers. Its capability to learn long-term dependencies allows to take advantage of the globality while preserving the locality obtained from the preceding convolutional layers (Guo et al. 2022). Moreover, since each individual has different visual characteristics, we inject speaker identity to the conformer encoder through an embedding table. As validated in our experiments, the speaker embedding helps the model to extract discriminative visual features, enhancing naturalness and intelligibility of the synthesised speech.

Variance Decoder

To ease the one-to-many mapping problem in LTS, the variance decoder aims to generate acoustic representation with rich variance information. As shown in Figure 1(c), the variance decoder consists of variance predictors and a conformer decoder. The variance predictors are composed of linguistic, pitch, and energy predictor, each of which aims to condition the corresponding variance information into the hidden visual representations $\boldsymbol{h}_v$. During training, we take the ground truth variance information to the hidden sequence, and use predicted value during inference. The following conformer decoder then converts the empowered hidden visual features to intermediate acoustic representations. In the following paragraphs, we describe the details of each predictor.

Linguistic Predictor   The presence of homophenes hinders the synthesis of intelligible speech with accurate pronunciation (Ephrat and Peleg 2017). Although the previous work (Kim, Hong, and Ro 2023) attempts to address the ambiguity of homophenes by leveraging text supervision, it requires manually annotated text labels and fails to enjoy the

benefits of the LTS system. To generate intelligible speech while preserving the self-supervised nature, we propose a linguistic predictor that disambiguates homophenes without the need for text labels.

To this end, we adopt quantised self-supervised speech representations. We extract continuous linguistic representations from raw waveforms using pre-trained SSL speech model, namely HuBERT (Hsu et al. 2021), and quantise the continuous representations for robust training[2]. Previous studies (Polyak et al. 2021; Lakhotia et al. 2021; Kreuk et al. 2022) report that the quantised speech representations from the specific layer of HuBERT contain elaborate linguistic features relevant to accurate pronunciation. We investigate the effects of different configurations of linguistic feature extraction, and empirically find that the representations from the 12th layer of HuBERT-LARGE[3], quantised by K-means algorithm with 200 clusters, exhibits the highest correlation with linguistic information.

Given the visual hidden features $\boldsymbol{h}_v$, the linguistic predictor aims to estimate the cluster indices of each frame. The linguistic predictor is optimised by cross-entropy (CE) loss which can be formulated as follows:

$$\mathcal{L}_l = \sum_{i=1}^{T_v} \text{CE}(\boldsymbol{l}_i, LinguisticPredictor(\boldsymbol{h}_{v,i})), \quad (1)$$

where $\boldsymbol{l}_i$ and $\boldsymbol{h}_{v,i}$ are the ith cluster index of the target sequence and the ith visual representation, respectively.

Pitch Predictor   Pitch plays an important role in synthesising realistic speech with natural prosody (Yasuda et al. 2019;

---

[2]Before quantisation, the continuous features are downsampled to match the length of video by nearest-neighbor interpolation.

[3]https://huggingface.co/facebook/hubert-large-ls960-ft

Łańcucki 2021). However, the pitch exhibits multiple variations across gender, age, and emotions, exacerbating the one-to-many problem in LTS. To accurately capture pitch information from lip motions, we construct a pitch predictor (Łańcucki 2021) that estimates the pitch sequence based on the hidden visual features.

Following (Łańcucki 2021), we extract the ground truth pitch values from the ground truth audio through pYIN algorithm (Mauch and Dixon 2014), and standardise them to have zero mean and unit variance for better sampling. The extracted pitch values are successively downsampled to match the temporal dimension of the visual features. The pitch predictor is optimised with L1 loss between the ground truth and predicted pitch sequence as follows:

$$\mathcal{L}_p = \sum_{i=1}^{T_v} ||\boldsymbol{p}_i - PitchPredictor(\boldsymbol{h}_{v,i})||_1, \quad (2)$$

where $\boldsymbol{p}_i$ is the ground truth pitch for the ith video frame.

Energy Predictor Energy represents the intensity of speech, which affects the volume and prosody of speech (Bulut, Lee, and Narayanan 2007). We obtain the target energy sequence by taking the L2 norm of the mel-spectrogram along the frequency-axis (Choi et al. 2021). To estimate the energy sequence from $\boldsymbol{h}_v$, we construct energy predictor (Ren et al. 2021), which is optimised by L1 loss between the ground truth and predicted energy sequence which can be formulated as follows:

$$\mathcal{L}_e = \sum_{i=1}^{T_v} ||\boldsymbol{e}_i - EnergyPredictor(\boldsymbol{h}_{v,i})||_1, \quad (3)$$

where $\boldsymbol{e}_i$ refers to the ith target energy value.

Each variance information is encoded into variance embeddings either through an embedding table (linguistic), or a single 1D convolution layer (pitch and energy). The variance embeddings are added to the visual representations $\boldsymbol{h}_v$, and the adapted representation is upsampled to match the time resolution of target mel-spectrogram. Lastly, the conformer decoder converts the adapted representations to a coarse mel-sepctrogram. We apply L1 loss between the ground truth mel-spectrogram and the predicted mel-spectrogram:

$$\mathcal{L}_{mel} = \sum_{j=1}^{T_m} ||\boldsymbol{Y}_j - \hat{\boldsymbol{Y}}_j||_1, \quad (4)$$

where Y_j denotes the jth frame of the ground truth mel-spectrogram with length T_m, and $\hat{Y}_j$ stands for the jth frame of predicted mel-spectrogram.

Note that the variance predictors simplify the acoustic target distribution by providing conditional information, thereby mitigating the one-to-many mapping issue (Ren et al. 2022). We analyse the effect of variance information in our experiment section.

Post-Net

Natural human speech comes with dynamic variations. However, simple reconstruction loss (L1 or L2 loss) is limited to capture such details, resulting in a blurry and over-smoothed synthetic speech (Liu et al. 2022). To further improve the sample quality, we apply a flow based post-net (Ren, Liu, and Zhao 2021) which elaborates the coarse-grained mel-spectrogram into a fine-grained one.

The architecture of the post-net is depicted in Figure 1(d). In training stage, the post-net transforms mel-spectrogram training data $\boldsymbol{x}$ into a tractable prior distribution through a series of invertible functions $\mathbf{f} = \mathbf{f}_0 \circ \mathbf{f}_1 \circ \cdots \mathbf{f}_k$, conditioned with the input and output of the conformer decoder, and the speaker embedding. The post-net is optimised with minimising the negative likelihood of data $\boldsymbol{x}$ as follows:

$$\log p_\theta(\boldsymbol{x}) = \log p_\theta(\boldsymbol{z}) + \sum_{i=1}^{k} \log |\det(\boldsymbol{J}(\mathbf{f}_i(\boldsymbol{x})))|, \quad (5)$$

$$\mathcal{L}_{post} = -\log p_\theta(\boldsymbol{x}), \quad (6)$$

where $p_\theta(\boldsymbol{z})$ is the tractable prior (isotropic multivariate Gaussian) over latent variable $\boldsymbol{z}$, and $\boldsymbol{J}$ is the Jacobian. During inference stage, we take samples $\boldsymbol{z}$ from the prior distribution and feed them into the post-net reversely to generate the final mel-spectrogram. As proved in (Ren et al. 2022), this flow-based module enhances the capability of modelling complex data distributions, which helps to address one-to-many mapping problem.

To summarise, the final loss ($\mathcal{L}_{final}$) is given by:

$$\mathcal{L}_{final} = \mathcal{L}_{mel} + \lambda_{var}\mathcal{L}_{var} + \lambda_{post}\mathcal{L}_{post}, \quad (7)$$

where $\mathcal{L}_{var} = \mathcal{L}_l + \mathcal{L}_p + \mathcal{L}_e$. In our experiments, we set $\lambda_{var} = \lambda_{post} = 0.1$.

Neural Vocoder

To convert the predicted mel-spectrogram into an audible waveform, we utilise pre-trained Fre-GAN (Kim et al. 2021) as our neural vocoder. Based on adversarial networks, it produces frequency-consistent waveform by adopting discrete wavelet transform (Daubechies 1988) as a lossless down-sampling method. Compared to Griffin-Lim algorithm (Griffin and Lim 1984) which is mostly adopted in previous works (Kim, Hong, and Ro 2021; He et al. 2022; Kim, Hong, and Ro 2023), the Fre-GAN vocoder shows exceptional performance in reconstructing waveforms. The vocoder is used only at inference process.

Experimental Settting

Datasets

GRID (Cooke et al. 2006) is one of the established datasets for lip to speech synthesis in a constrained environment. It contains 33 speakers and 50 minutes of short video clips per speaker. The number of vocabularies in GRID is only 51 and the speakers always face forward with nearly no head movement. In our experiment, the dataset is split into train (80%), validation (10%), and test set (10%) by sampling equally from all speakers.

Lip2Wav (Prajwal et al. 2020) is a large-scale benchmark dataset for an unconstrained and large vocabulary lip to speech synthesis. It comprises real-world lecture recordings featuring 5 distinct speakers, with about 20 hours of video for each speaker. Our experiment is conducted on

2 speakers, Lip2Wav-Chemistry and Lip2Wav-Chess, as in GlowLTS (He et al. 2022). The two speakers are trained jointly in a multi-speaker setting, and both are equally divided into 80-10-10% for train, validation, and test sets.

Preprocessing

Audio data are resampled to 16kHz and transformed to mel-spectrograms with 40ms window length, 10ms hop length, and 80 mel filterbanks. To achieve the temporal synchronisation with the audio, the video data are resampled to 25 frames per second, resulting in a fixed ratio of 1 to 4 between the lengths of video and audio. We then extract 68 face landmarks for each video frame using FaceAlignment[4]. Based on the landmarks, the lip regions are aligned to a fixed position, and cropped to their centers with a dimension of 112×112. The cropped images are converted to grayscale.

Model Configuration

We construct the identical model architecture for each dataset, with the exception of the conformer encoder. For the constrained GRID dataset, the conformer encoder is designed with 6 attention heads and a hidden dimension of 384, and for the unconstrained Lip2Wav, the encoder is designed with 8 heads and a hidden dimension of 512. We follow the recent works (Mira et al. 2022; Kim, Hong, and Ro 2023) for the configuration of 3D convolution and ResNet18. The speaker embedding is obtained from an embedding table, where each speaker identity is converted into a fixed size embedding vector. The pitch and energy predictors are composed of two 1D convolution layers (Łańcucki 2021), while the linguistic predictor consisted of four 1D convolutions.

Training Details

Our model is trained on four NVIDIA A5000 GPUs with a batch size of 64. We use AdamW optimiser (Loshchilov and Hutter 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. The learning rate is fixed to 2×10^{-4}, and the weight decay is set to 10^{-6}. For training the GRID dataset, we randomly sample consecutive sequence with a length of 50, and the model is trained for 400 epochs. For the Lip2Wav dataset, we sample contiguous 75 frames and train our model for 900 epochs. To prevent overfitting, we apply data augmentations: horizontal flipping with probability of 50%, and random masking with fixed position throughout all frames. The masked area is randomly sampled within the range from 10×10 to 30×30.

Evaluation Metrics

The performance of the proposed method is evaluated with both qualitative and quantitative evaluation metrics. For qualitative evaluation, we conduct mean opinion score (MOS) test, wherein 30 domain-expert speakers assess the quality of 30 random speech clips for naturalness and intelligibility on a scale of 1 to 5. Naturalness of speech represents how close the speech is to that of human utterance. Intelligibility focuses solely on the successful delivery of linguistic contents; high scores are given if one can clearly

identify the contents even if it sounds unnatural. Moreover, we also compute word error rate (WER) and character error rate (CER) of 300 random test samples for quantitative evaluation. For error calculation, we obtain the transcriptions of speech clips by using publicly available automatic speech recognition (ASR) model (Radford et al. 2023) pre-trained on 438k hours of English corpus.

Based on the above evaluation metrics, we compare our model against the ground truth, vocoded, and generated samples from recent LTS models which show promising results: VCA-GAN (Kim, Hong, and Ro 2021), SVTS (Mira et al. 2022), and MT (Kim, Hong, and Ro 2023). We follow the official implementation for VCA-GAN[5] and MT[6]. For SVTS, since there is no official implementation, we reproduce it based on the official code of MT which takes SVTS as its backbone network. For a fair comparison, all the LTS models are trained on the same settings, and the predicted mel-spectrograms are converted to audible speech by Fre-GAN vocoder (Kim et al. 2021).

Experimental Results

We evaluate our method in qualitative and quantitative manner, and investigate each variance prediction pipeline. Further, the effectiveness of each component in our model is verified by the ablation study. The evaluations underscore a notable enhancement over the other LTS models, and each module provides distinct contribution to increasing the quality of synthesised speech.

Qualitative Results

To evaluate the quality of the generated speech, we conduct MOS for naturalness and intelligibility, and the results are presented in Table 1. For the GRID dataset, our proposed method achieves the highest naturalness and intelligibility among all generated samples. Especially, the generated speech of the proposed method closely approximates the vocoded quality with a minor gap of 0.28 in naturalness and 0.13 in intelligibility. For Lip2Wav, the overall generation quality degrades due to the unconstrained nature of the dataset. However, the proposed method produces speech with promising quality, outperforming the existing methods by a significant margin. Note that MT (Kim, Hong, and Ro 2023) is not applicable to Lip2Wav experiment, since it requires text information for training while the dataset does not provide text labels.

Moreover, to intuitively demonstrate the effectiveness of the proposed method, we conduct mel-spectrogram visualisation analysis. In Figure 2, we depict the generated mel-spectrograms along with the ground truth and vocoded mel-spectrogram. Especially in red boxes, our model produces a detailed and sharp mel-spectrogram with distinct harmonics, showing close resemblance to the ground truth mel-spectrogram. However, other methods ((c)-(e)) suffer from blurry and over-smoothed results. This indicates that our

---

[4]https://github.com/ladrianb/face-alignment

[5]https://github.com/ms-dot-k/Visual-Context-Attentional-GAN

[6]https://github.com/ms-dot-k/Lip-to-Speech-Synthesis-in-the-Wild

Method	GRID				Lip2Wav			
	Nat. ↑	Intel. ↑	WER ↓	CER ↓	Nat. ↑	Intel. ↑	WER ↓	CER ↓
Ground Truth	4.82 ± 0.04	4.82 ± 0.05	12.20	7.03	4.80 ± 0.05	4.78 ± 0.05	4.12	2.58
Vocoded	4.74 ± 0.05	4.79 ± 0.05	12.44	7.18	4.63 ± 0.06	4.73 ± 0.06	6.05	4.19
VCA-GAN	3.46 ± 0.07	4.10 ± 0.08	17.62	9.55	2.05 ± 0.08	2.71 ± 0.10	48.73	32.51
SVTS	3.35 ± 0.09	3.97 ± 0.09	23.30	13.12	1.77 ± 0.07	2.18 ± 0.10	61.09	41.01
MT	2.42 ± 0.09	3.08 ± 0.12	30.56	18.14	N/A	N/A	N/A	N/A
Ours	$\mathbf{4.46 \pm 0.07}$	$\mathbf{4.63 \pm 0.07}$	**17.07**	**9.17**	$\mathbf{4.15 \pm 0.08}$	$\mathbf{3.69 \pm 0.10}$	**34.71**	**22.57**

Table 1: Evaluation results. MOS results are presented with 95% confidence interval. 'Nat.' and 'Intel.' represent MOS for naturalness and intelligibility, respectively. Note that MT (Kim, Hong, and Ro 2023) cannot be trained on the Lip2Wav dataset since the model requires text transcription to be trained. ↑ denotes higher is better, ↓ denotes lower is better.

(a) Ground Truth   (b) Vocoded   (c) VCA-GAN
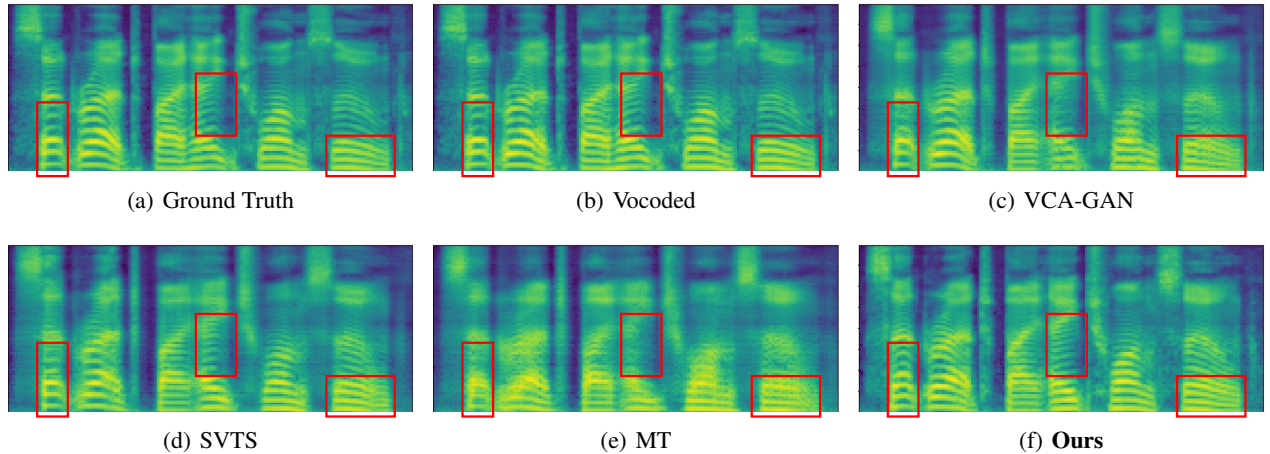
(d) SVTS   (e) MT   (f) **Ours**

Figure 2: Visualisation of mel-spectrogram. Note that the proposed method better captures fine details with frequency correlations compared to other methods ((c)-(e)), particularly in red boxes.

method can effectively learn the complex one-to-many mapping function, which consequently lead to natural and intelligible synthetic speech.

Quantitative Results

As a quantitative evaluation, we compare the WER and CER of the synthesised speech with those of the ground truth and vocoded speech. For the GRID dataset, the error rates are obtained by directly comparing the ASR transcriptions with the provided ground truth texts. For the Lip2Wav dataset, since the dataset does not provide text labels, we manually annotate the ground truth texts and compare them with the ASR transcription results.

The results are shown in Table 1. The proposed model clearly shows the lowest WER and CER on both GRID and Lip2Wav datasets. This demonstrates our method can synthesise highly intelligible speech by effectively reducing the homophene problems. Despite the higher error rates in the Lip2Wav dataset compared to GRID dataset, the proposed method achieves significantly lower WER and CER compared to all the other models, making a gap larger than 10% point. This explicitly supports that the proposed model is readily applicable to the unconstrained environments.

Analysis on Acoustic Variance Information

To verify the effectiveness of the acoustic variance conditions, we examine the similarities between 300 pairs of the synthesised and ground truth speech. For pitch, we compute the moments of pitch distribution (mean (μ), standard deviation (σ), skewness (γ), and kurtosis (κ)) and analyse on how much the values resemble those of the ground truth. The results are shown in Table 2. Each of the four values from the output of the proposed model stands the closest to those from the ground truth speech, especially the skewness value deviating only by 0.042. This demonstrates that our model can generate speech with highly accurate pitch contour. With the absence of the pitch predictor, the kurtosis shows a slight deviance. However, closing the gap of standard deviation with minimum changes in mean, skewness and kurtosis clearly supports that the pitch predictor explicitly contributes to producing high-quality results.

For energy, we calculate the frame-wise mean absolute error (MAE) between energy of generated speech and that of ground truth speech. As shown in Table 3, the MAE from the proposed model reports the lowest among all other models by a distinct margin. This implies that the energy of the generated speech from the proposed model closely resembles that from the ground truth speech. The influence of the

Method	μ	σ	γ	κ
Ground Truth	77.90	101.84	0.696	−1.217
VCA-GAN	102.14	95.98	0.124	−1.454
SVTS	98.85	97.95	0.198	−1.518
MT	83.42	91.82	0.450	−1.347
Ours	**79.74**	**101.63**	**0.654**	−1.284
w/o pitch	79.81	102.39	0.652	−**1.282**

Table 2: Mean (μ), standard deviation (σ), skewness(γ), and kurtosis(κ) of the pitch distribution for ground truth and synthesised audio.

Method	VCA-GAN	SVTS	MT	**Ours**	**Ours** *w/o* E.
MAE ↓	4.155	4.275	5.314	**3.886**	3.959

Table 3: The MAE between the energy of ground truth and that of synthesised audio. "E." stands for energy.

energy predictor is also confirmed by the increase of MAE when the predictor is removed.

Analysis on Self-Supervised Features

The significance of SSL speech models is proven by the recent studies (Baevski et al. 2020; Hsu et al. 2021), and the research is further explored with the utilisation of intermediate representations on various downstream tasks (wen Yang et al. 2021; Chang, Yang, and Lee 2022). Outputs from the first layer are used to extract speaker identity in (Fan et al. 2021; Choi et al. 2021), and (Lakhotia et al. 2021; Lee et al. 2022) utilise the middle layer to obtain linguistic representation. Particularly, (Lakhotia et al. 2021) reports that the number of K-means clusters clearly affects the model performance when using the quantised linguistic representation.

To find the optimal linguistic feature configuration for our model, we compute WER, CER, and phoneme error rate (PER) on the GRID validation set using various feature extraction settings. To be specific, linguistic features are obtained from 1st, 12th, and 24th layer outputs of HuBERT. The continuous linguistic features are then quantised with 100, 200 clusters, and the cluster indices are used as targets for the linguistic predictor. Table 4 demonstrates that the outputs from 12th layer of HuBERT quantised with 200 clusters produce the most intelligible speech, achieving the lowest PER and CER. While the same configuration but 100 clusters achieves the lowest WER, larger number of clusters shows lower PER and CER. Considering the fact that phoneme accuracy is closely related to the accurate pronunciation (Zhang et al. 2022), the configuration with the lowest PER generates the most intelligible speech.

Ablation Study

To verify the effect of each module in the proposed method, we conduct an ablation study on the Lip2Wav dataset using 7-scale comparative MOS (CMOS), WER, and CER. In CMOS, 30 domain experts listen to the audio samples from two systems and compare the quality from -3 to +3.

#clusters	layer	WER ↓	PER ↓	CER ↓
100	1	18.02	9.58	10.04
100	12	**16.53**	10.77	11.39
100	24	17.57	8.92	10.10
200	1	17.62	9.73	9.76
200	12	17.12	**8.91**	**9.70**
200	24	29.17	9.59	16.03

Table 4: Evaluation on different configurations of linguistic feature extraction. #clusters denotes the number of K-means clusters and layer means the layer index of HuBERT.

Method	Nat. ↑	Intel. ↑	WER ↓	CER ↓
Ours	0.00	0.00	34.71	22.57
w/o linguistic	−0.90	−0.70	42.51	27.99
w/o pitch	−1.06	−0.61	39.96	26.30
w/o energy	−0.42	−0.62	40.58	26.46
w/o post-net	−1.48	−0.57	40.05	25.48
w/o e_{spk}	−0.48	−0.56	42.33	27.04

Table 5: CMOS, WER, and CER results of an ablation study.

As shown in Table 5, the results of the ablation study clearly support that each component independently contributes to improving the quality of the synthetic speech. Notably, the absence of the linguistic predictor results in the largest quality degradation in speech intelligibility, WER, and CER. This proves the effectiveness of the linguistic predictor in clarifying homophenes, which connects to speech generation with accurate pronunciation. The significance of the acoustic variance information, especially pitch, is validated by the quality degradation when such information is not considered. Removing the post-net shows the largest decrease in naturalness, highlighting the effectiveness of the module in producing fine details of acoustic features. The importance of speaker information e_{spk} is proven by the degraded quality when the information is excluded.

Conclusion

In this paper, we propose a novel LTS system that generates speech close to human-level quality in both naturalness and intelligibility. We directly tackle the inherent one-to-many mapping problems of LTS, and address them by providing linguistic and acoustic variance information. We further refine the generated speech by enhancing modelling capability. Both qualitative and quantitative experiments clearly demonstrate that the proposed method improves the overall quality of the synthesised speech, outperforming the previous works by a notable margin. We also verify the effectiveness of each proposed component through the ablation study, and analyse the effect of the variance information from various perspectives. For the future work, we will continue to enhance the generated speech quality by adopting audio-visual SSL models. We also aim to simplify the overall generation pipeline with the inclusion of neural vocoder, making a fully end-to-end architecture.

Broader Impact

By employing the proposed LTS system, numerous positive societal impacts can be realised, including the dubbing of silent videos and the simulation of natural utterances for individuals with speech impairments. However, alongside these advantages, there exist potential threats associated with the misuse of our system, such as the generation of fake speech and voice phishing. Furthermore, as the LTS system enables one to comprehend conversations from a distance, there is a risk of its use in invading personal privacy.

Acknowledgements

References

Baevski, A.; Hsu, W.-N.; Conneau, A.; and Auli, M. 2021. Unsupervised Speech Recognition. In *NeurIPS*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. Wav2Vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *NeurIPS*.

Black, A. W.; Zen, H.; and Tokuda, K. 2007. Statistical Parametric Speech Synthesis. In *Proc. ICASSP*.

Bulut, M.; Lee, S.; and Narayanan, S. S. 2007. Analysis of Emotional Speech Prosody in Terms of Part of Speech Tags. In *Proc. Interspeech*.

Chang, H.-J.; Yang, S.-w.; and Lee, H.-y. 2022. DistilHuBERT: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit BERT. In *Proc. ICASSP*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proc. ICML*.

Choi, H.-S.; Lee, J.; Kim, W.; Lee, J.; Heo, H.; and Lee, K. 2021. Neural Analysis and Synthesis: Reconstructing Speech from Self-Supervised Representations. In *NeurIPS*.

Chung, J. S.; Senior, A.; Vinyals, O.; and Zisserman, A. 2017. Lip Reading Sentences in the Wild. In *Proc. CVPR*.

Cooke, M.; Barker, J.; Cunningham, S.; and Shao, X. 2006. An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition. *The Journal of the Acoustical Society of America*, 120(5): 2421–2424.

Daubechies, I. 1988. Orthonormal Bases of Compactly Supported Wavelets. *Communications on pure and applied mathematics*, 41(7): 909–996.

Elias, I.; Zen, H.; Shen, J.; Zhang, Y.; Jia, Y.; Weiss, R. J.; and Wu, Y. 2021. Parallel Tacotron: Non-Autoregressive and Controllable TTS. In *Proc. ICASSP*.

Ephrat, A.; and Peleg, S. 2017. Vid2Speech: Speech Reconstruction from Silent Video. In *Proc. ICASSP*.

Fan, Z.; Li, M.; Zhou, S.; and Xu, B. 2021. Exploring Wav2Vec 2.0 on Speaker Verification and Language Identification. In *Proc. Interspeech*.

Griffin, D.; and Lim, J. 1984. Signal Estimation from Modified Short-Time Fourier Transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2): 236–243.

Gulati, A.; Qin, J.; Chiu, C.-C.; Parmar, N.; Zhang, Y.; Yu, J.; Han, W.; Wang, S.; Zhang, Z.; Wu, Y.; and Pang, R. 2020. Conformer: Convolution-Augmented Transformer for Speech Recognition. In *Proc. Interspeech*.

Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; and Xu, C. 2022. CMT: Convolutional Neural Networks Meet Vision Transformers. In *Proc. CVPR*.

He, J.; Zhao, Z.; Ren, Y.; Liu, J.; Huai, B.; and Yuan, N. 2022. Flow-Based Unconstrained Lip to Speech Generation. In *Proc. AAAI*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proc. CVPR*.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 29: 3451–3460.

Hunt, A. J.; and Black, A. W. 1996. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database. In *Proc. ICASSP*.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2012. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1): 221–231.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL*.

Kim, C.; Um, S. Y.; Yoon, H.; and Kang, H. G. 2022. FluentTTS: Text-Dependent Fine-Grained Style Control for Multi-Style TTS. In *Proc. Interspeech*.

Kim, J.-H.; Lee, S.-H.; Lee, J.-H.; and Lee, S.-W. 2021. Fre-GAN: Adversarial Frequency-Consistent Audio Synthesis. In *Proc. Interspeech*.

Kim, M.; Hong, J.; and Ro, Y. M. 2021. Lip to Speech Synthesis with Visual Context Attentional GAN. In *NeurIPS*.

Kim, M.; Hong, J.; and Ro, Y. M. 2023. Lip-to-Speech Synthesis in the Wild with Multi-Task Learning. In *Proc. ICASSP*.

Kim, M.; Yeo, J. H.; and Ro, Y. M. 2022. Distinguishing Homophenes Using Multi-Head Visual-Audio Memory for Lip Reading. In *Proc. AAAI*.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In *NeurIPS*.

Kreuk, F.; Polyak, A.; Copet, J.; Kharitonov, E.; Nguyen, T.-A.; Rivière, M.; Hsu, W.-N.; Mohamed, A.; Dupoux, E.; and Adi, Y. 2022. Textless Speech Emotion Conversion using Discrete and Decomposed Representations. In *Proc. EMNLP*.

Kumar, Y.; Jain, R.; Salik, K. M.; Shah, R. R.; Yin, Y.; and Zimmermann, R. 2019. Lipper: Synthesizing Thy Speech Using Multi-View Lipreading. In *Proc. AAAI*.

Lakhotia, K.; Kharitonov, E.; Hsu, W.-N.; Adi, Y.; Polyak, A.; Bolte, B.; Nguyen, T.-A.; Copet, J.; Baevski, A.; Mohamed, A.; et al. 2021. On Generative Spoken Language Modeling from Raw Audio. *Transactions of the Association for Computational Linguistics*, 9: 1336–1354.

Łańcucki, A. 2021. Fastpitch: Parallel Text-to-Speech with Pitch Prediction. In *Proc. ICASSP*.

Le Cornu, T.; and Milner, B. 2015. Reconstructing Intelligible Audio Speech from Visual Speech Features. In *Proc. Interspeech*.

Le Cornu, T.; and Milner, B. 2017. Generating Intelligible Audio Speech from Visual Speech. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 25(9): 1751–1761.

Lee, S.-H.; Kim, J.-H.; Chung, H.; and Lee, S.-W. 2021a. VoiceMixer: Adversarial Voice Style Mixup. In *NeurIPS*.

Lee, S.-H.; Kim, S.-B.; Lee, J.-H.; Song, E.; Hwang, M.-J.; and Lee, S.-W. 2022. HierSpeech: Bridging the Gap between Text and Speech by Hierarchical Variational Inference using Self-supervised Representations for Speech Synthesis. In *NeurIPS*.

Lee, S.-H.; Yoon, H.-W.; Noh, H.-R.; Kim, J.-H.; and Lee, S.-W. 2021b. Multi-SpectroGAN: High-Diversity and High-Fidelity Spectrogram Generation with Adversarial Style Combination for Speech Synthesis. In *Proc. AAAI*.

Lin, Y.; Gou, Y.; Liu, Z.; Li, B.; Lv, J.; and Peng, X. 2021. Completer: Incomplete Multi-view Clustering via Contrastive Prediction. In *Proc. CVPR*.

Liu, J.; Li, C.; Ren, Y.; Chen, F.; and Zhao, Z. 2022. DiffSinger: Singing Voice Synthesis via Shallow Diffusion Mechanism. In *Proc. AAAI*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Proc. ICLR*.

Ma, P.; Petridis, S.; and Pantic, M. 2022. Visual Speech Recognition for Multiple Languages in the Wild. *Nature Machine Intelligence*, 4(11): 930–939.

Mauch, M.; and Dixon, S. 2014. pYIN: A Fundamental Frequency Estimator Using Probabilistic Threshold Distributions. In *Proc. ICASSP*.

Midgley, N. 2006. New Technology Catches Hitler Off Guard. *Telegraph*.

Mira, R.; Haliassos, A.; Petridis, S.; Schuller, B. W.; and Pantic, M. 2022. SVTS: Scalable Video-to-Speech Synthesis. In *Proc. Interspeech*.

Petridis, S.; Stafylakis, T.; Ma, P.; Cai, F.; Tzimiropoulos, G.; and Pantic, M. 2018. End-to-End Audiovisual Speech Recognition. In *Proc. ICASSP*.

Polyak, A.; Adi, Y.; Copet, J.; Kharitonov, E.; Lakhotia, K.; Hsu, W.-N.; Mohamed, A.; and Dupoux, E. 2021. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech*.

Popov, V.; Vovk, I.; Gogoryan, V.; Sadekova, T.; and Kudinov, M. 2021. Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech. In *Proc. ICML*.

Prajwal, K.; Mukhopadhyay, R.; Namboodiri, V. P.; and Jawahar, C. 2020. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis. In *Proc. CVPR*.

Prenger, R.; Valle, R.; and Catanzaro, B. 2019. WaveGlow: A Flow-based Generative Network for Speech Synthesis. In *Proc. ICASSP*.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proc. ICML*.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2021. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *Proc. ICLR*.

Ren, Y.; Liu, J.; and Zhao, Z. 2021. PortaSpeech: Portable and High-Quality Generative Text-to-Speech. In *NeurIPS*.

Ren, Y.; Tan, X.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2022. Revisiting Over-Smoothness in Text to Speech. In *Proc. ACL*.

Shen, J.; Pang, R.; Weiss, R. J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. 2018. Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions. In *Proc. ICASSP*.

Skerry-Ryan, R.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.; Clark, R.; and Saurous, R. A. 2018. Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. In *Proc. ICML*.

Smith, J. 1987. When Lip Reading between the Lines Had the Subtitles Beat by a Long Sight. *LA Times*.

Sun, G.; Zhang, Y.; Weiss, R. J.; Cao, Y.; Zen, H.; and Wu, Y. 2020. Fully-Hierarchical Fine-grained Prosody Modeling for Interpretable Speech Synthesis. In *Proc. ICASSP*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is All You Need. In *NeurIPS*.

wen Yang, S.; Chi, P.-H.; Chuang, Y.-S.; Lai, C.-I. J.; Lakhotia, K.; Lin, Y. Y.; Liu, A. T.; Shi, J.; Chang, X.; Lin, G.-T.; Huang, T.-H.; Tseng, W.-C.; tik Lee, K.; Liu, D.-R.; Huang, Z.; Dong, S.; Li, S.-W.; Watanabe, S.; Mohamed, A.; and yi Lee, H. 2021. SUPERB: Speech Processing Universal Performance Benchmark. In *Proc. Interspeech*.

Yadav, R.; Sardana, A.; Namboodiri, V. P.; and Hegde, R. M. 2021. Speech Prediction in Silent Videos Using Variational Autoencoders. In *Proc. ICASSP*.

Yamamoto, R.; Song, E.; and Kim, J.-M. 2020. Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In *Proc. ICASSP*.

Yasuda, Y.; Wang, X.; Takaki, S.; and Yamagishi, J. 2019. Investigation of Enhanced Tacotron Text-to-Speech Synthesis Systems with Self-Attention for Pitch Accent Language. In *Proc. ICASSP*.

Zhang, G.; Song, K.; Tan, X.; Tan, D.; Yan, Y.; Liu, Y.; Wang, G.; Zhou, W.; Qin, T.; Lee, T.; et al. 2022. Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech. In *Proc. Interspeech*.