VCA-GAN (**?**), SVTS (**?**), and MT (**?**). 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 (**?**).

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 (**?**) 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 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.

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.

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 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 (**??**), and the research is further explored with the utilisation of intermediate representations on various

[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	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.

downstream tasks (**??**). Outputs from the first layer are used to extract speaker identity in (**??**), and (**??**) utilise the middle layer to obtain linguistic representation. Particularly, (**?**) 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 1^{st}, 12^{th}, and 24^{th} 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 12^{th} 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 (**?**), 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. 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 $\boldsymbol{e}_{spk}$ 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.

#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 $\boldsymbol{e}_{spk}$	−0.48	−0.56	42.33	27.04

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

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

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00212845, Multi-modal Speech Processing for Human-Computer Interaction). Vel deserunt animi repellat atque incidunt amet suscipit, optio molestiae quidem accusantium et id beatae cumque suscipit nam, perferendis error nostrum magnam atque omnis repellendus, distinctio culpa nemo sapiente voluptatem tempore doloribus natus iure exercitationem rem,