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,