HierSpeech: Bridging the Gap between Text and Speech by Hierarchical Variational Inference using Self-supervised Representations for Speech Synthesis

Sang-Hoon Lee[1] Seung-Bin Kim[2] Ji-Hyun Lee[2] Eunwoo Song[3] Min-Jae Hwang[3] Seong-Whan Lee[2]

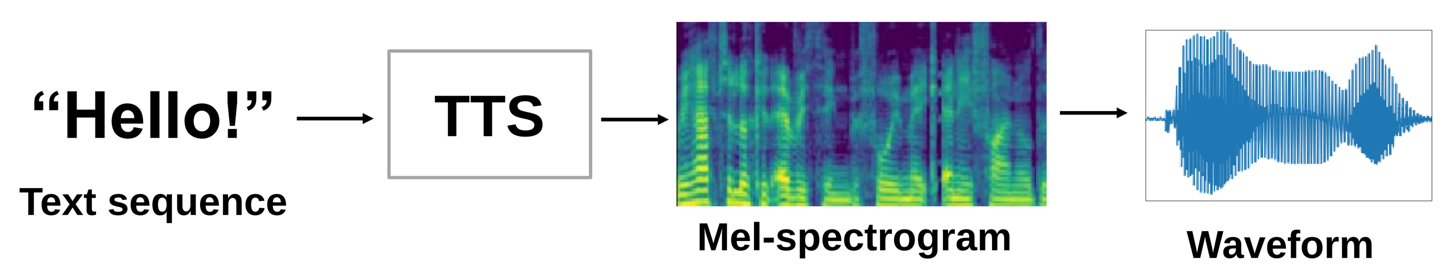
[1] Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea, sh_lee@korea.ac.kr
[2] Department of Artificial Intelligence, Korea University, Seoul, Korea, {sb-kim, jihyun-lee, sw.lee}@korea.ac.kr
[3] NAVR Corp., Seongnam, Korea, {eunwoo.song, min-jae.hwang}@navercorp.com

Objective

The fundamental objectives of our proposed method are

- Introducing an additional linguistic representation to bridge the gap between text and speech
- Adopting the hierarchical conditional variational autoencoder to connect linguistic and acoustic representations, and to lean each attribute hierarchically
- Proposing an untranscribed text-to-speech framework, which can adapt to a novel speaker by utilizing self-supervised speech representations without text transcripts

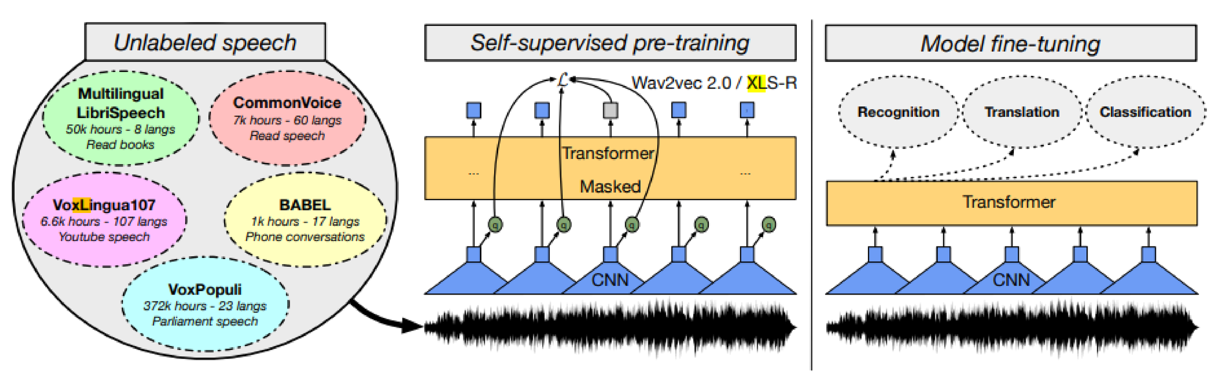
Text-to-Speech (TTS)



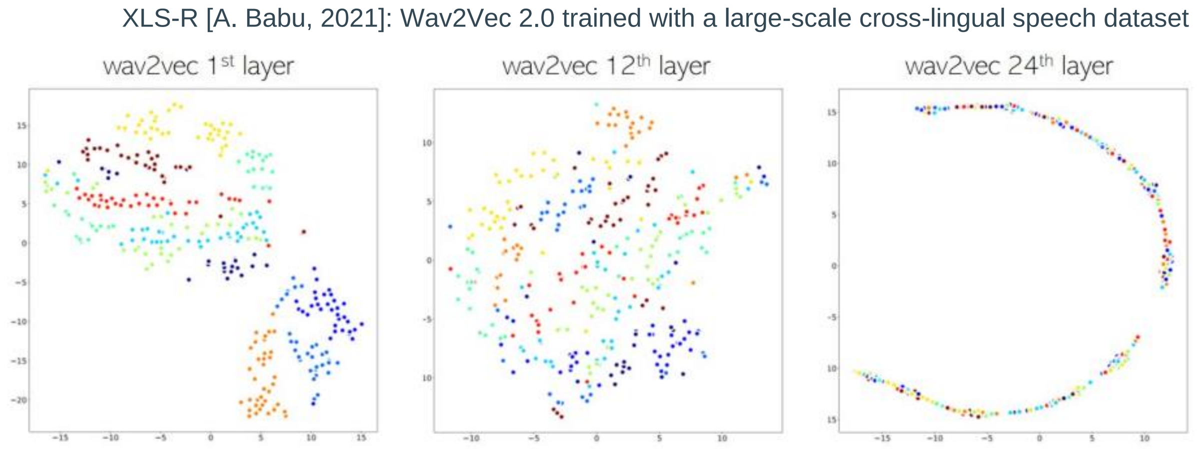
TTS pipeline

Self-supervised Speech Representation

- Self-supervised model can learn useful information from large-scale unlabeled data
- The representations from the middle layer of the pre-trained model contain rich linguistic information (e.g., the pronunciation of speech)
- Self-supervised model for speech: Wav2Vec 2.0 [A. Baevski, 2020]

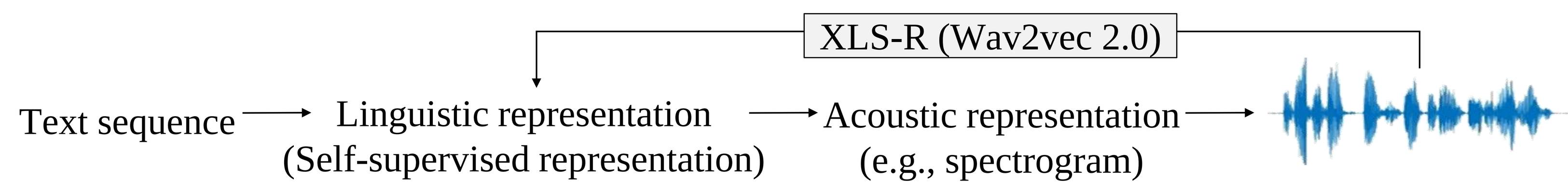


XLS-R [A. Babu, 2021]: Wav2Vec 2.0 trained with a large-scale cross-lingual speech dataset



Visualization of intermediate representations of Wav2vec 2.0 using t-SNE in NANSY [H.-S. Choi, 2021]

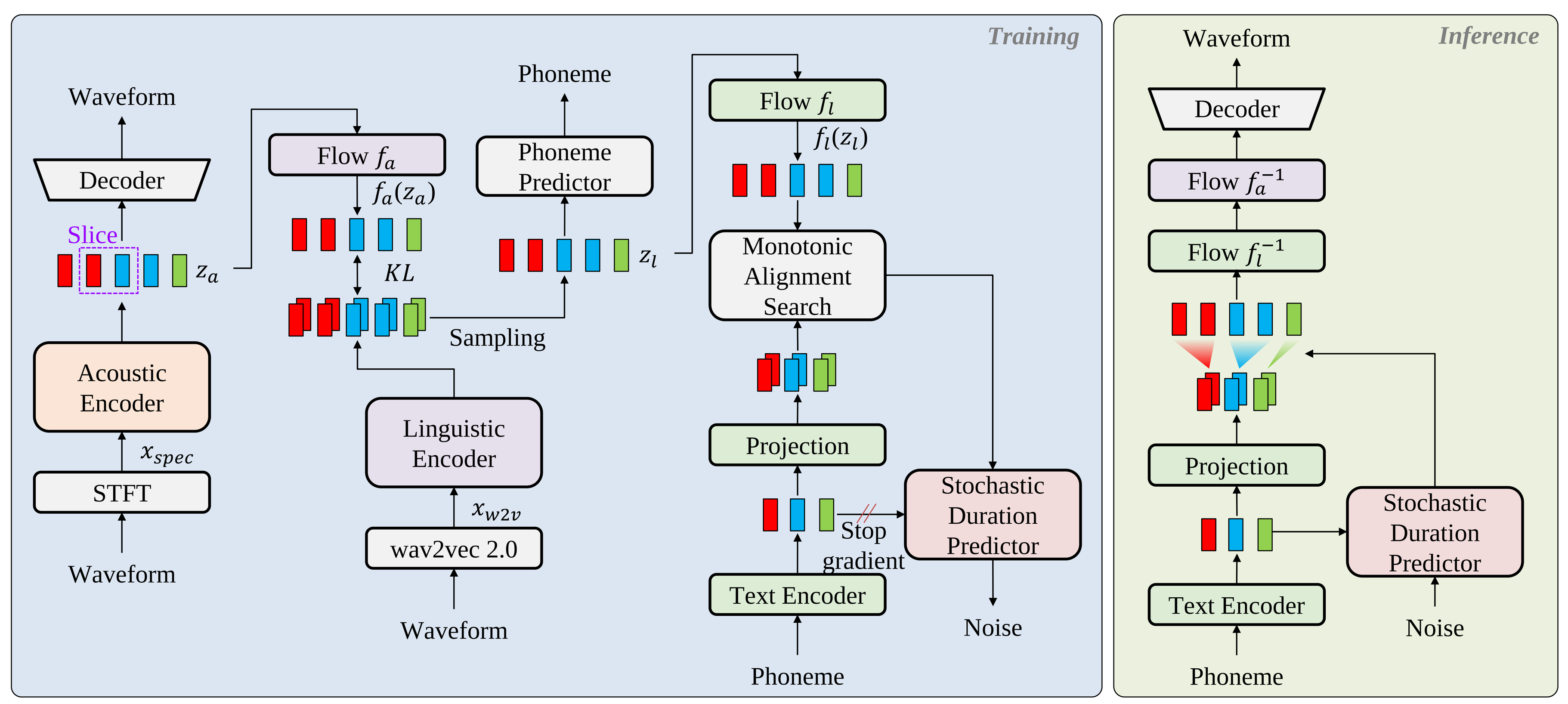
Hierarchical Text-to-Speech Pipeline



Hierarchical TTS pipeline

HierSpeech

- We integrate the linguistic representations learned by self-supervised learning (SSL) models into the end-to-end TTS pipeline
- Baseline TTS model: VITS [J. Kim, 2021]
- SSL model: XLS-R (Wav2Vec 2.0), we used the middle layer (12th) of XLS-R



- Model Architecture

- **Decoder**: Generate raw waveform audio from acoustic representation z_a
- **Acoustic Encoder**: Extract the acoustic representation z_a from linear spectrogram
- **Linguistic Encoder**: Extract the linguistic representation z_l from the wav2vec 2.0
- **Phoneme Predictor**: Enforce the linguistic characteristics in z_l
- **Text Encoder**: Extracting the linguistic prior distribution

- Hierarchical variational inference + Adversarial training

- **Reconstruction:** $l1$ distance of Mel-spectrogram between the GT and reconstructed waveform using STFT and Mel-scale transformation
- **KL Divergence between acoustic posterior and prior distribution**
- **KL Divergence between linguistic posterior and prior distribution**
- **Adversarial training:** GAN loss + Feature Matching loss

- HierSpeech-U: Untranscribed text-to-speech model

- Finetuning the model without text transcripts

Experiment and Result

<Audio Samples>

- Dataset

1. VCTK (46 hours for 108 speakers)
2. LibriTTS (110 hours for 1,151 speakers)

- Analysis of self-supervised representations

The representations from the middle layer of XLS-R contain a small amount of speaker information

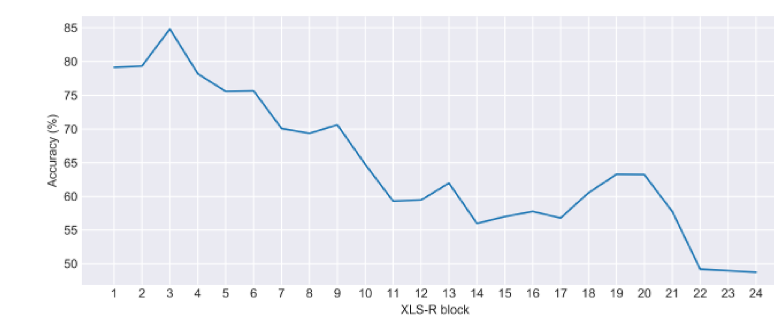


Figure 7: Speaker classification on self-supervised representations from different layers of XLS-R.

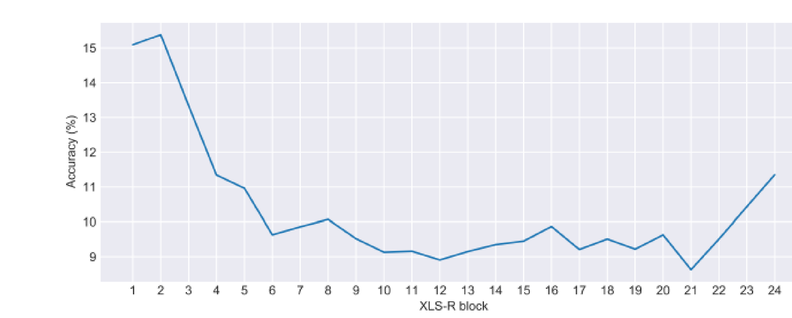


Figure 8: Speaker classification on linguistic representations from different layers of XLS-R.

- Evaluation

1. Subjective metrics: naturalness/similarity Mean Opinion Score (nMOS/sMOS)
2. Objective metrics: Phoneme Error Rate (PER), Word Error Rate (WER), Equal Error Rate (EER) of the Automatic Speaker Verification model, etc.

Table 3: The TTS evaluation results on the VCTK dataset.

Method	nMOS	sMOS	PER	WER	EER	MCD	RMSE$_{f0}$	DDUR	Speed (kHz)	Real-time
GT	4.06±0.02	3.34±0.03	5.64	18.94	4.03	-	-	-	-	-
GT (HiFi-GAN)	4.03±0.02	3.30±0.03	5.94	19.52	5.04	1.25	28.32	-	6,484.09	×294.06
Tacotron2	3.76±0.02	3.16±0.03	11.73	22.48	9.11	4.18	35.30	0.49	263.94	×11.97
Glow-TTS	3.95±0.02	3.09±0.03	11.77	26.40	5.33	4.31	32.98	0.38	1,410.75	×63.97
PortaSpeech	3.97±0.02	3.15±0.03	11.35	25.46	5.48	4.34	32.89	0.43	1,163.21	×52.75
VITS	4.02±0.02	3.19±0.03	9.16	25.54	3.83	4.27	32.93	0.37	**1,610.77**	×**72.83**
HierSpeech (Ours)	**4.04±0.02**	**3.22±0.03**	**5.78**	**19.55**	**3.74**	**4.05**	**32.15**	**0.33**	1,459.95	×66.21

Table 4: The speaker transfer evaluation results on the LibriTTS dataset.

Method	nMOS	sMOS	PER	WER	EER	MCD	RMSE$_{f0}$	DDUR	Speed (kHz)	Real-time
GT	4.04±0.03	3.40±0.03	7.01	18.28	4.45	-	-	-	-	-
VITS	3.96±0.03	3.26±0.03	13.62	29.83	5.00	**4.37**	34.18	1.09	**1,781.40**	×**80.78**
HierSpeech (Ours)	**3.98±0.03**	3.26±0.03	**7.47**	**20.34**	5.00	4.42	**32.95**	**0.72**	1,678.79	×76.13

- Untranscribed TTS

Table 6: Results for untranscribed text-to-speech. We compare few-shot speaker adaptation performance of HierSpeech-U with that of HierSpeech. Both models use the pre-trained HierSpeech which is trained using VCTK and LibriTTS datasets. We used 10 unseen speakers of VCTK dataset as novel speakers, and fine-tuned each model with 20 samples from each speaker.

Method	Transcript	nMOS	sMOS	PER	WER	EER	MCD	RMSE$_{f0}$	DDUR
GT	-	4.13±0.10	3.38±0.10	4.26	16.69	4.14	-	-	-
HierSpeech	✓	4.09±0.10	3.18±0.11	4.40	16.95	6.40	3.96	29.56	0.28
HierSpeech-U	✗	4.08±0.09	3.15±0.12	3.71	15.85	6.40	4.09	30.64	0.36

References

[J. Kim, 2021] J. Kim et al., "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," *ICML*, 2021.
[H.-S. Choi, 2021] H.-S. Choi et al., "Neural Analysis and Synthesis Reconstructing Speech from Self-supervised Representations," *NeurIPS*, 2021.
[A. Baevski, 2020] A. Baevski et al., "Wav2vec 2.0: A Framework for Self-supervised Learning of Speech Representation," *NeurIPS*, 2020.
[A. Babu, 2021] A. Babu, et al., "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.