P4-12-TTS (#184)

Disentangling the Prosody and Semantic Information with Pre-trained Model for In-Context Learning based Zero-Shot Voice Conversion

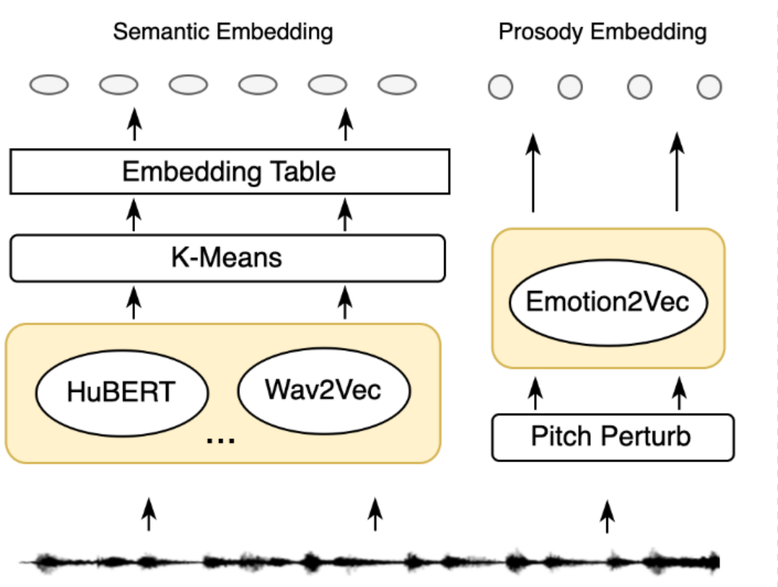
Zhengyang Chen[1], Shuai Wang[2,3], Mingyang Zhang[2,3], Xuechen Liu[4], Junichi Yamagishi[4], Yanmin Qian[1, †]

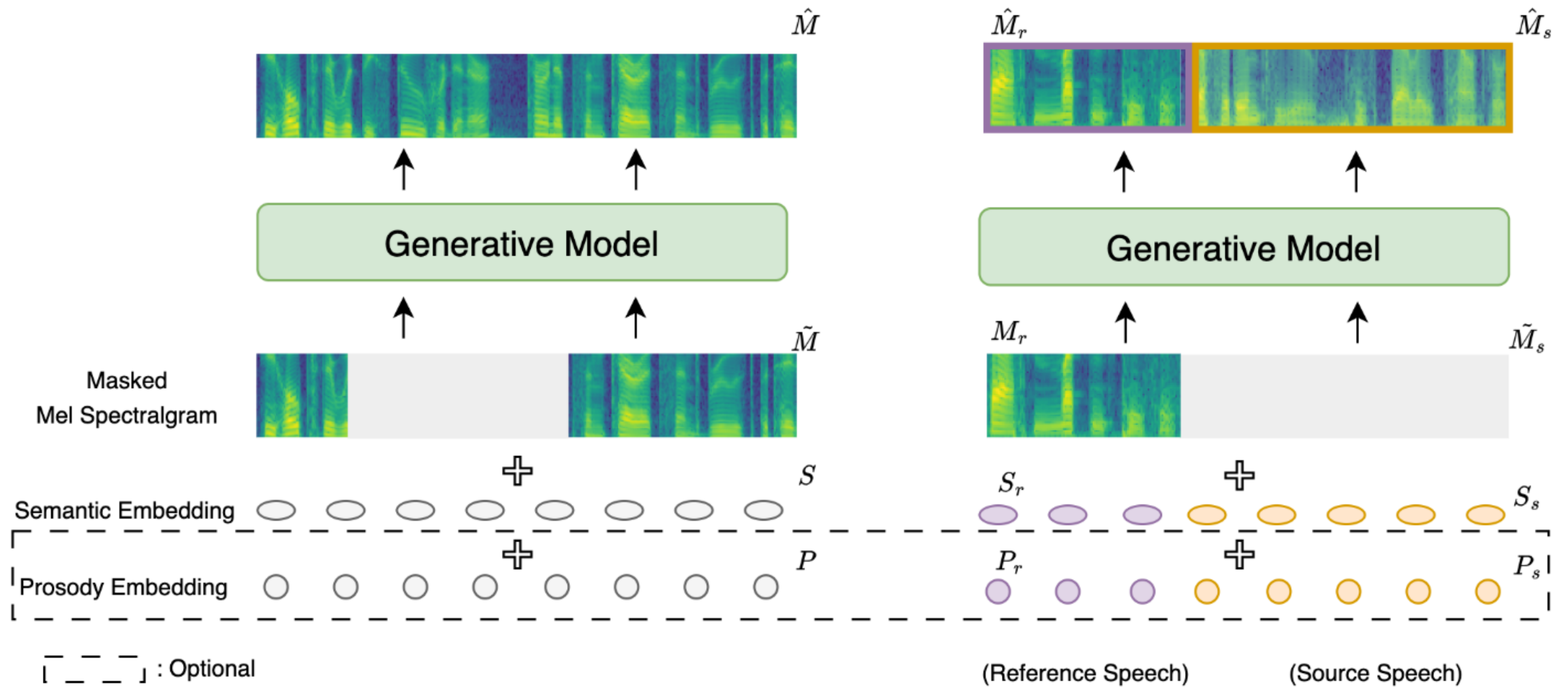
[1]AudioCC Lab, CS Dept, Shanghai Jiao Tong University, Shanghai, China

[2]Shenzhen Research Institute of Big Data, Shenzhen, China [3]Chinese University of Hong Kong, Shenzhen, China

[4]National Institute of Informatics, Tokyo, Japan

System Overview





- ❑ K-means is applied for all kinds of pre-trained models to extract semantic tokens.
- ❑ Speech is perturbed and then fed into the Emotion2vec model to extract prosody embedding.
- ❑ Flow-matching algorithm is leveraged in our generative model.

Voice Conversion Task Evaluation

Table 1. Voice Conversion Results on LibriTTS dataset. Prosody embedding is not used for ICL-VC system's input in this evaluation.

System	SECS	CER	Natualness MOS
YourTTS	0.824	3.83	3.86 ± 0.13
RefXVC	0.797	2.22	4.02 ± 0.12
ICL-VC-HuBERT	0.846	2.66	4.33 ± 0.11
ICL-VC-Wav2Vec	0.845	2.23	4.37 ± 0.12
ICL-VC-Wav2Vec-XLSR	0.856	5.09	4.32 ± 0.11
Ground Truth	-	1.69	4.51 ± 0.09

- ❑ System is trained on LibriTTS train-clean-360 and tested on LibriTTS test-clean.
- ❑ K-means tokenization is effective even the pre-trained model (wav2vec) is not trained with k-means related objectives.
- ❑ Enlarging the speaker coverage of pre-trained model (XLSR) can improve voice conversion task's speaker similarity.

Ablation Study on Reference Speech Duration

- ❑ Longer reference speech allows the system to better model the speaker's timbre information.
- ❑ Our ICL-VC system achieves high cosine speaker similarity even with a very short (3-second) reference utterance.

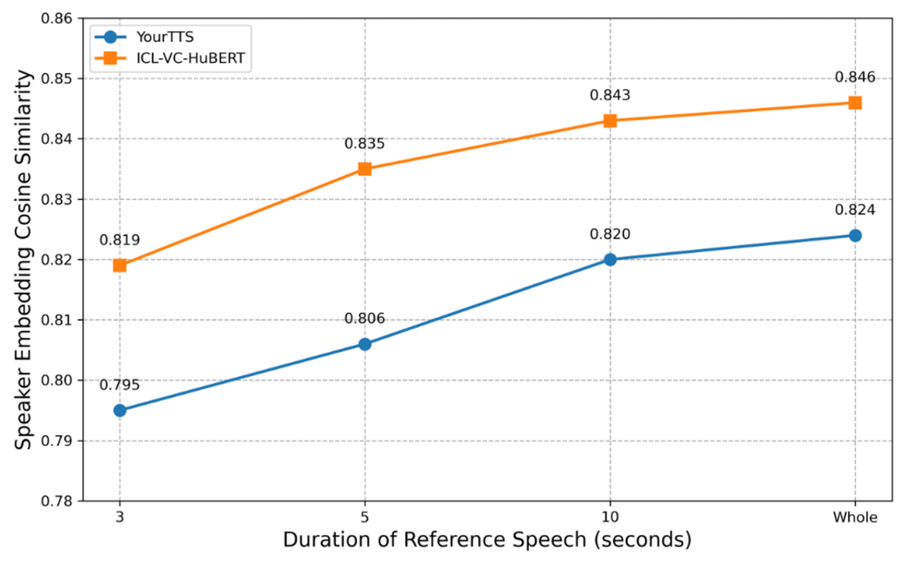


Fig. 2. The relationship between the speaker embedding cosine similarity (SECS) and reference speech duration.

Prosody Preserving Ability Evaluation about Voice Conversion Task

Table 2. Voice Conversion Results on ESD dataset. In this experiment, only the semantic tokens extracted from HuBERT pre-trained model are used in our ICL-VC system. The Pitch Corr and Energy Corr correspond to the Pearson correlation introduced in section 3.4.

System	SECS	CER	Pitch Corr	Energy Corr	Prosody MOS	Naturalness MOS
YourTTS	0.706	10.68	0.683	0.898	4.02 ± 0.22	3.33 ± 0.15
RefXVC	0.748	16.50	0.512	0.695	3.79 ± 0.13	3.53 ± 0.12
ICL-VC	0.800	5.39	0.608	0.867	3.34 ± 0.18	3.77 ± 0.17
+ Pitch & Energy	0.769	5.31	0.727	0.935	4.12 ± 0.17	3.61 ± 0.16
+ Prosody Emb	0.789	4.33	0.671	0.900	4.19 ± 0.16	4.14 ± 0.13
Ground Truth	-	3.21	-	-	-	4.41 ± 0.08

- ❑ System is trained on LibriTTS train-clean-360 and tested on Emotion Speech Database.
- ❑ Replacing the prosody embedding with pitch & energy token can effectively preserve prosody, but degrade the naturalness of speech.
- ❑ Leveraging the prosody embedding can simultaneously preserve prosody and achieve natural speech.

Full Paper *Demo Page*