Latent linguistic embedding for cross-lingual text-to-speech and voice conversion

Hieu-Thi Luong, Junichi Yamagishi

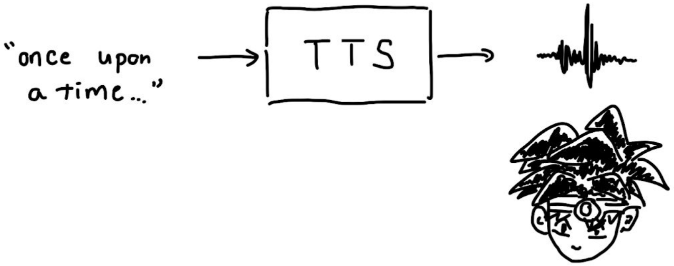
Inter-University Research Institute Corporation / Research Organization of Information and Systems

National Institute of Informatics

System **T07** for both intra-lingual and cross-lingual tasks of the Voice Conversion Challenge 2020

Text-to-speech (TTS)

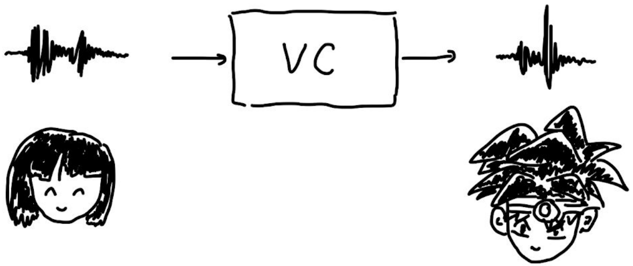
Generating speech with voice of a target speaker from a given text input.



Applications: audio books, computer screen reader, machine-human communications,...

Voice conversion (VC)

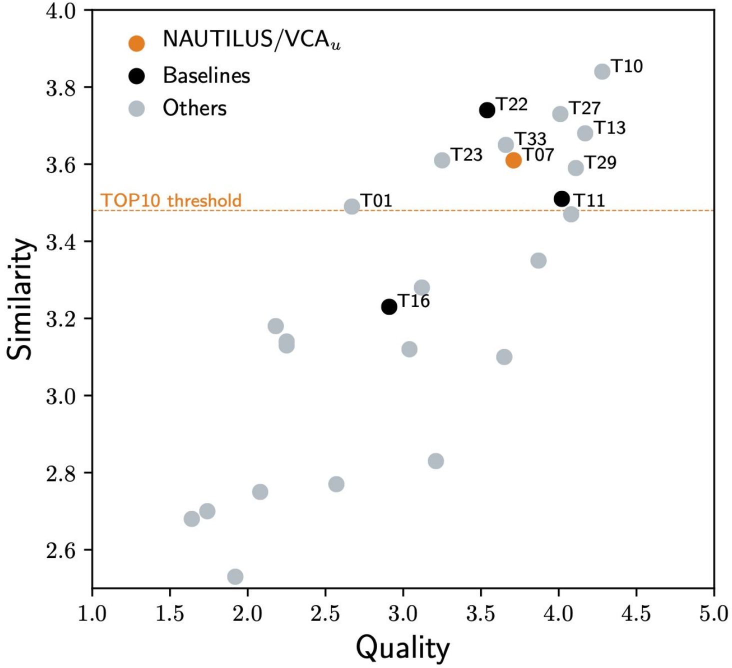
Changing voice of a speech utterance to that of a target while maintaining linguistic content.



Applications: movie dubbing, voice imitation for entertainment industry, voice avatar (for social media or video games),...

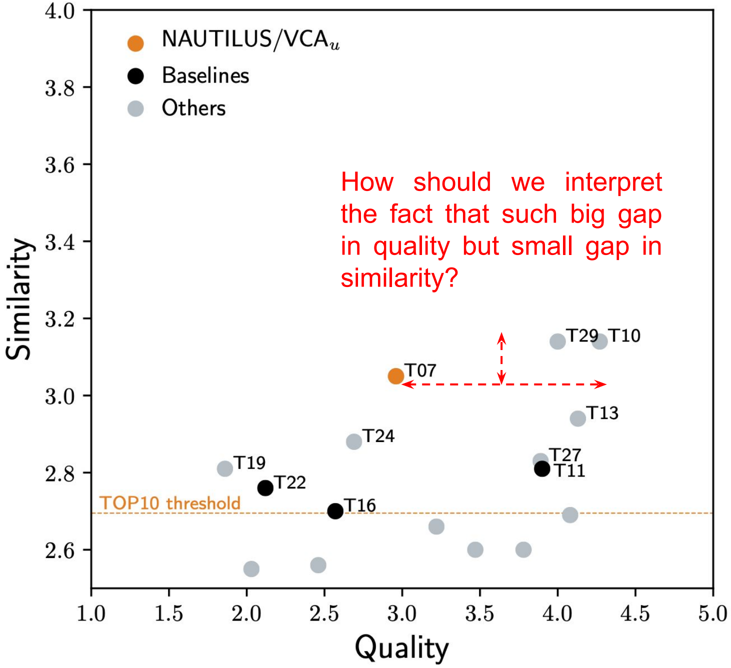
Cross-lingual speech generation is the scenario in which speech utterances are generated with the voices of target speakers in the language not spoken by them originally

VCC2020 Results



Task 1: intra-lingual VC

four English speakers



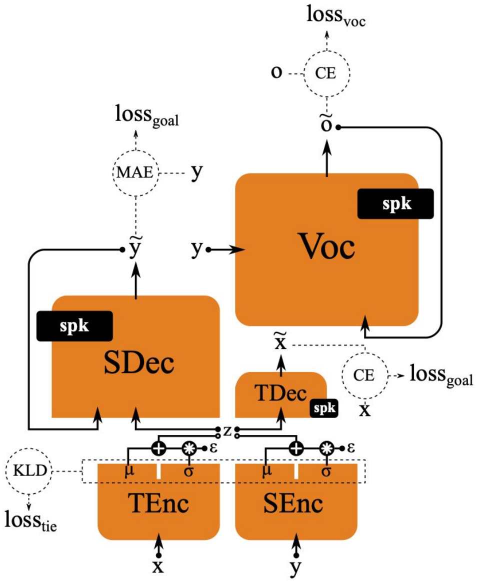
How should we interpret the fact that such big gap in quality but small gap in similarity?

Task 2: cross-lingual VC

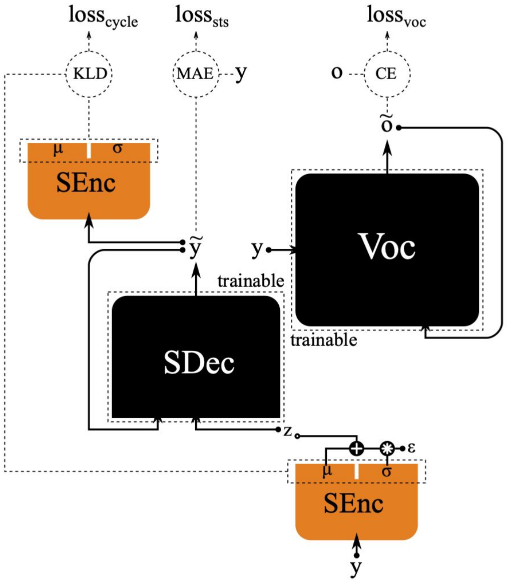
two Finnish, two German, and two Mandarin speakers

NAUTILUS$^{\text{cross-lingual}}$

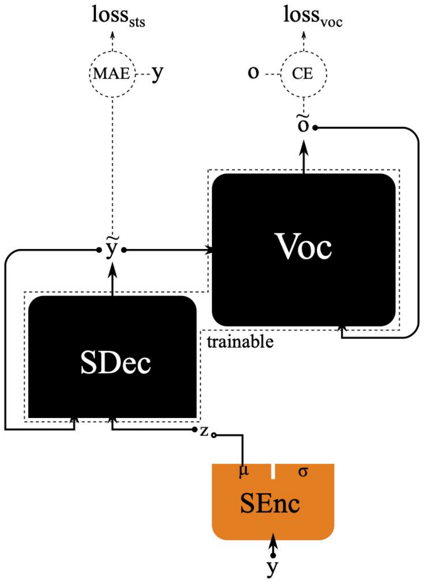
Cross-language speaker adaptation for a unified cross-lingual TTS/VC speech generation system.



Initial training: jointly train the text-speech multimodal system to obtain a robust **English** latent linguistic embedding



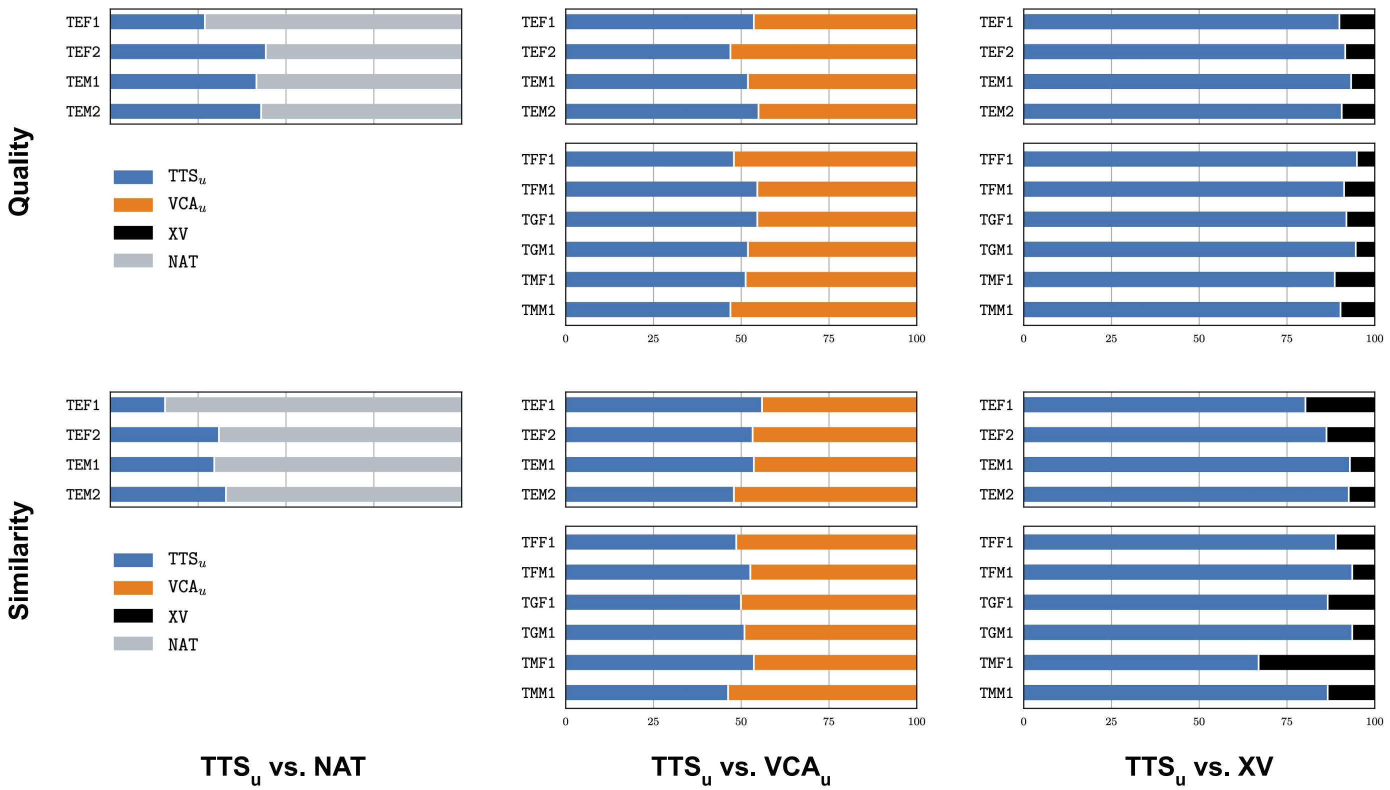
Step 1 - Adaptation: tuning with speech data of **foreign** speakers



Step 2 - Welding: jointly tune to increase the compatable

Cross-lingual TTS/VC

The proposed cross-lingual TTS and VC systems maintain a consistency performance between the two modes.



Conclusion

The **NAUTILUS** system has the ability to peform cross-language speaker adaptation for both TTS and VC interfaces.

The generated speech has an accented-like characteristic but more research is needed to confirm this observation.

Relevant materials can be found at www.hieuthi.com

Target speaker (TFF1)	“During the following years he tried unsuccessfully to get it into production”	SEF2_E30001.wav 🔊
🔊	TTS 🔊	VC 🔊

Nautilus noun

/'nɔtɚləs/