SSW11 Speech Synthesis Workshop
Aug 26-28 2021

Preliminary study on using vector quantization latent spaces for TTS/VC systems with consistent performance

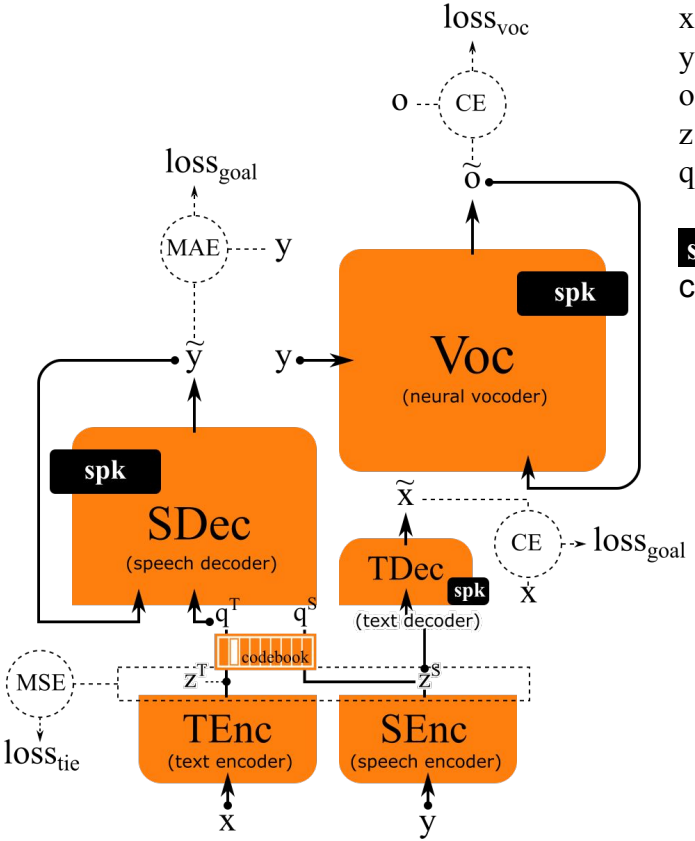
Hieu-Thi Luong, Junichi Yamagishi
National Institute of Informatics, JAPAN

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

Motivation

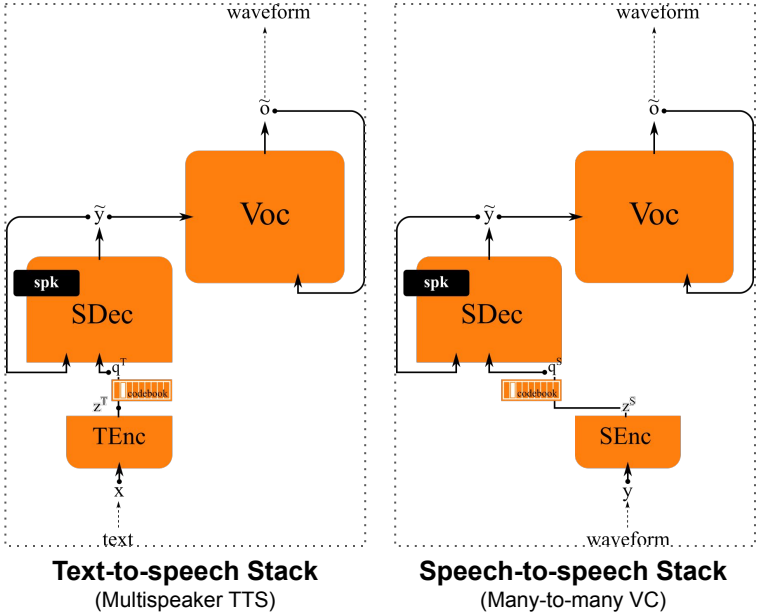
- In general, speech synthesis research and development (TTS: Text-to-speech, VC: Voice conversion,...) focus on the **end product** that is the **natural** and **expressive** synthesized speech (with or without an emphasis on **speaker individuality**).
- However, by training latent representations with specific traits, we opens up the possibilities to use in many applications/scenarios. Specifically, a **discrete latent feature** is a low bit-rate representation comparing with **continuous one**.
- In this work, we investigated the use of **vector quantization variational autoencoder** (VQVAE) component to model the linguistic latent space of our **unified TTS/VC voice cloning system** (name: NAUTILUS).
- Beside general evaluations on subjective quality/similarity, we also show how our carefully designed system can be used to **investigate the latent space**, which is a preliminary study for future research on this topic.

NAUTILUS-VQ



MAE: mean-absolute error
MSE: mean-square error
CE: cross-entropy
x: phonetic feature
y: acoustic feature
o: waveform
z: LLE
q: quantized LLE

spk: with speaker components (ohv)



Text-to-speech Stack
(Multispeaker TTS)

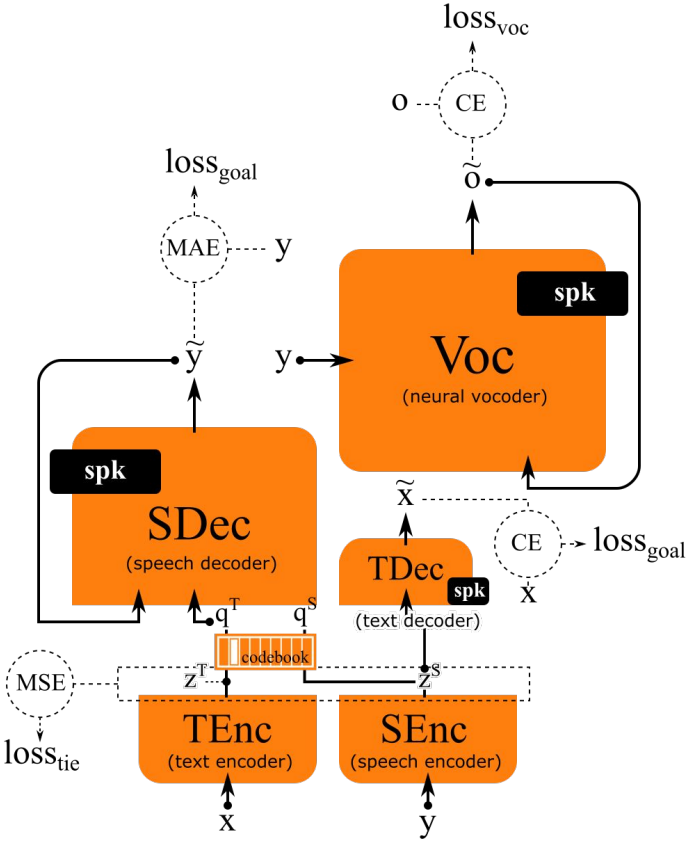
Speech-to-speech Stack
(Many-to-many VC)

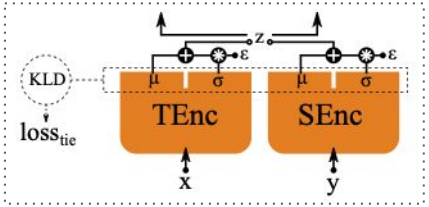
(←left) **A voice cloning system** that is a combination of TTS and VC (NAUTILUS*) with a **vector quantization** (VQ) component used to turn the continuous latent linguistic embedding (LLE) into discrete vector.

**We proposed NAUTILUS system with continuous LLE in our previous publication:*

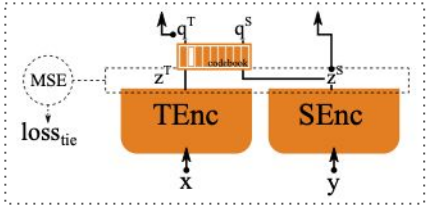
[1] Luong, Hieu-Thi, and Junichi Yamagishi. "Nautilus: a versatile voice cloning system." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2967-2981.

NAUTILUS-VQ: 1.INITIAL TRAINING





VAE-based encoders
(Original)



VQVAE-based encoders
(Proposed)

The main difference with the original is that we replace **VAE-based** encoders with **VQVAE-based** encoders which change the setup for the initial **supervised** (text+speech) training steps:

$$loss_{train} = loss^{tts}_{train} + \alpha_{sts} loss^{sts}_{train} + \alpha_{stt} loss^{stt} + \beta\, loss_{tie} \quad (1)$$

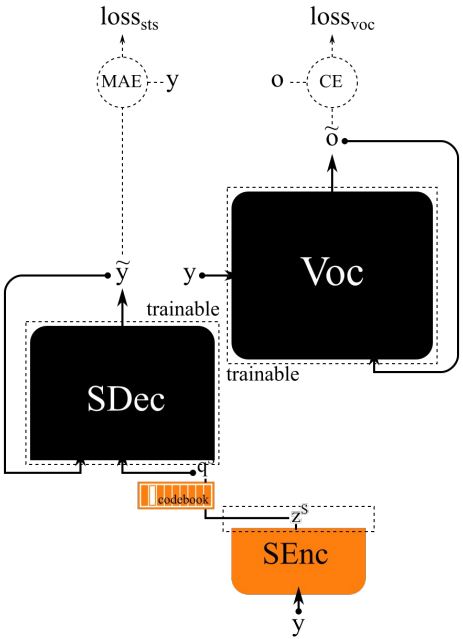
$$loss^{tts}_{train} = loss_{tts} + \delta_{VQ} loss^{T}_{VQ} + \delta_{C} loss^{T}_{C} \quad (2)$$

$$loss^{sts}_{train} = loss_{sts} + \delta_{VQ} loss^{S}_{VQ} + \delta_{C} loss^{S}_{C} \quad (3)$$

$$loss_{tie} = \left|\left| sg(z^T) - z^S \right|\right|^2_2 \quad (4)$$

Equation (1) is the jointly-trained of a **multimodal neural net** with joint-goal and tied-layer losses [1]
(Interchangeable encoder modules)
Equation (2) and (3) take on a form of a typical **VQVAE** setup.

NAUTILUS-VQ: 2.VOICE CLONING



Step 1 - Adaptation

Using pretrained modules from previous stage, we want to "**clone**" new voices of unseen speakers using their untranscribed speech data:

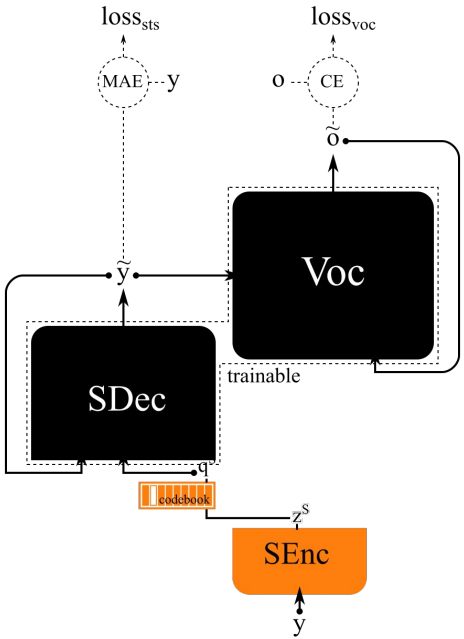
Step 1(←left): Adapting **speech encoder** and **neural vocoder** independently using STS stack (no text involved)

$$loss_{adapt} = loss_{sts} \quad (5)$$ (tuning speech decoder)

$$loss'_{adapt} = loss_{voc} \quad (6)$$ (tuning neural vocoder)

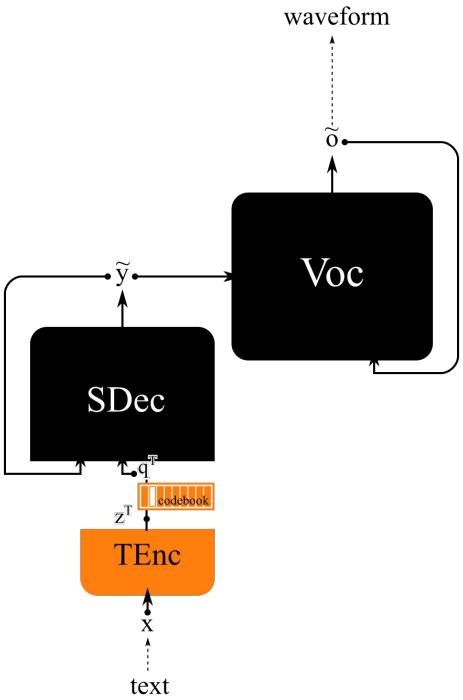
Step 2(right→): Tuning **speech decoder** and **neural vocoder** together using STS stack:

$$loss_{weld} = loss_{sts} + \gamma loss_{voc} \quad (7)$$



Step 2 - "Welding"

NAUTILUS-VQ: 3.TTS/VC INFERENCE

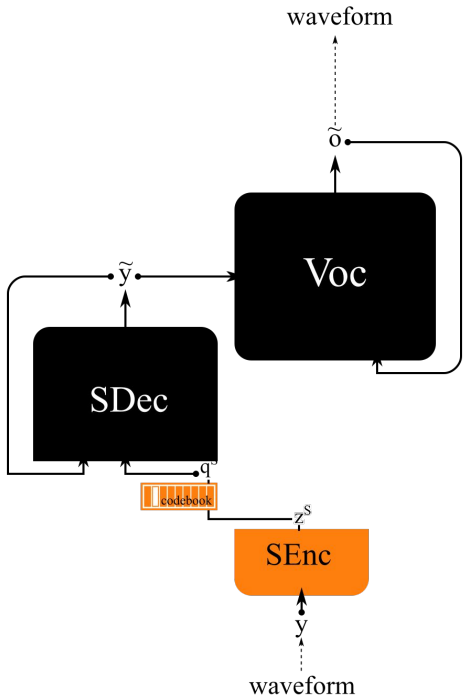


Inference/Text-to-Speech

Even though we only used speech encoder in the voice cloning stage, we can use either text or speech encoder to create a completed TTS/VC system

→ This is the case as we believe in the **consistency** between the text and speech encoders which we have deployed various policies (**jointly train TTS/VC**, **"tie" latent spaces**,...) in initial training stage to enforce it.

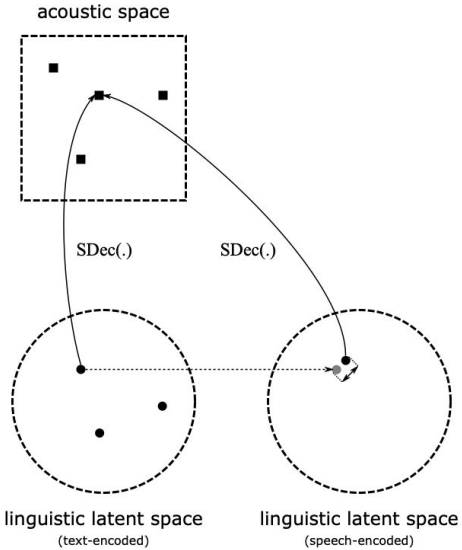
*figures are highly simplified for presentation purposes, the network architecture and details setup can be found in our paper and in [1]



Inference/Voice Conversion

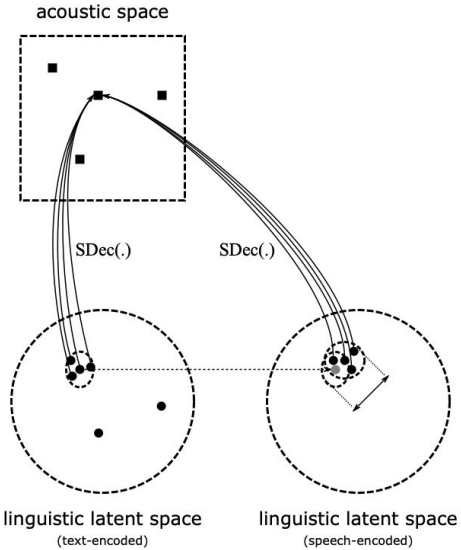
Shaping the linguistic latent spaces

Consistency? (Ideally) Text and speech of the same content will be transformed to an identical LLE sequence by text and speech encoders



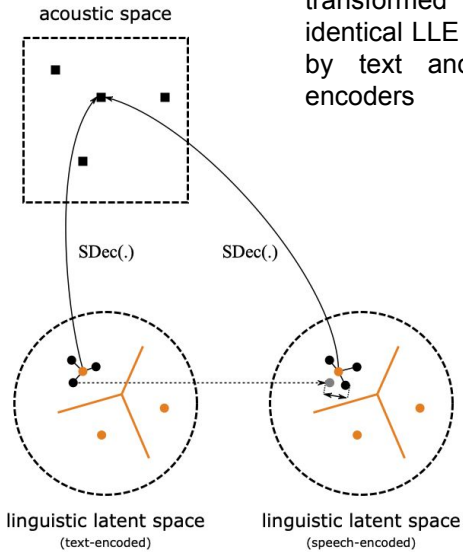
(a) Standard

Simple, but text and spech is not **one-to-one** relation.



(b) VAE-based [1]

Denoise training, **density-wise** tying function, has weaknesses of VAE



(c) VQVAE-based

Discrete latent feature, ???

The purpose of the multimodal setup is to train **highly consistent** and **interchangeable text encoder** and **speech encoder**. We enforce this by assuming certain properties about the linguistic latent space.

Experiments

Japanese voice cloning TTS/VC system using untranscribed speech of target speakers:

- **Initial training**: jointly train (supervised) using **speech** and **text** from ~236 hours of the 16kHz CSJ corpus, and ~134 hours of a 24kHz in-house corpus.
- **Voice cloning**: using only **speech** of target speakers (Table 1) to clone new voices for the TTS/VC system.

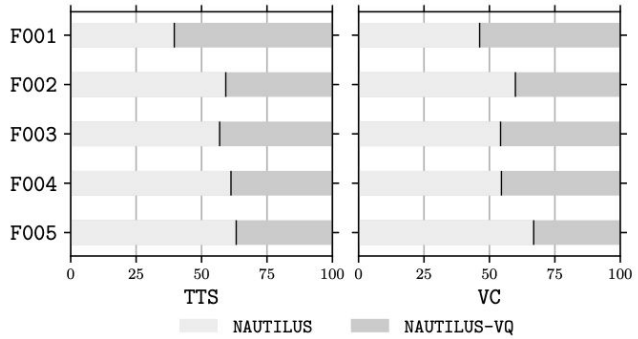
**Details experiment setups and network architecture can be found in the paper.*

Table 1: Japanese target speakers for voice cloning task.

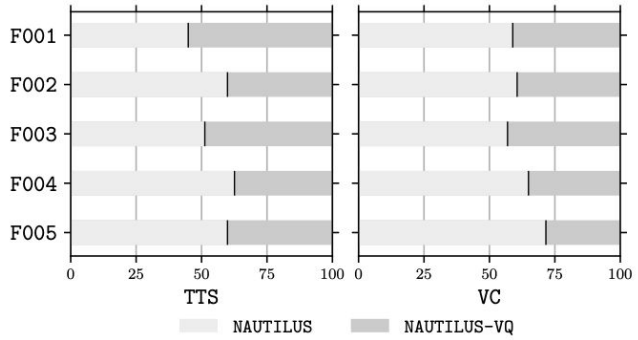
Speaker	Gender	Quantity	Duration
F001	female	483 utt.	45.0 min
F002	female	481 utt.	44.4 min
F003	female	484 utt.	47.4 min
F004	female	468 utt.	40.8 min
F005	female	485 utt.	47.6 min
XL10	female	10 utt.	55 s
		125 utt.	10.9 min
		500 utt.	44.5 min
		2000 utt.	2.9 h
		8750 utt.	12.9 h

Evaluations

samples



Quality



Similarity

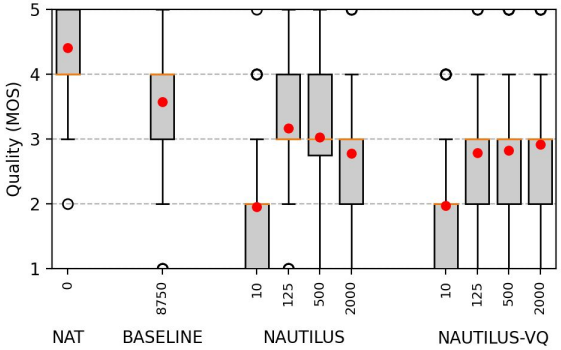
Exp01 (←left): comparing **NAUTILUS** and **NAUTILUS-VQ** on voice cloning task with ~45 minutes of untranscribed speech.

→ ***NAUTILUS*** *is better in most data points.*

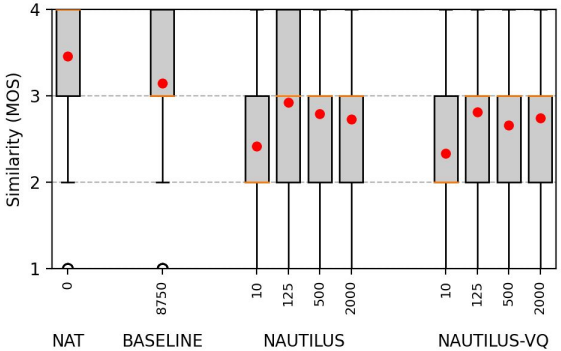
Exp02 (right→): comparing proposed voice cloning systems with a conventional SD TTS system (**BASELINE**) trained with ~12 hours of transcribed speech of a speaker.

→ ***NAUTILUS*** *and* ***NAUTILUS-VQ*** *are not as good as* ***BASELINE*** *but using less demanding data.*

Each speaker/system/task was evaluated 300 times.



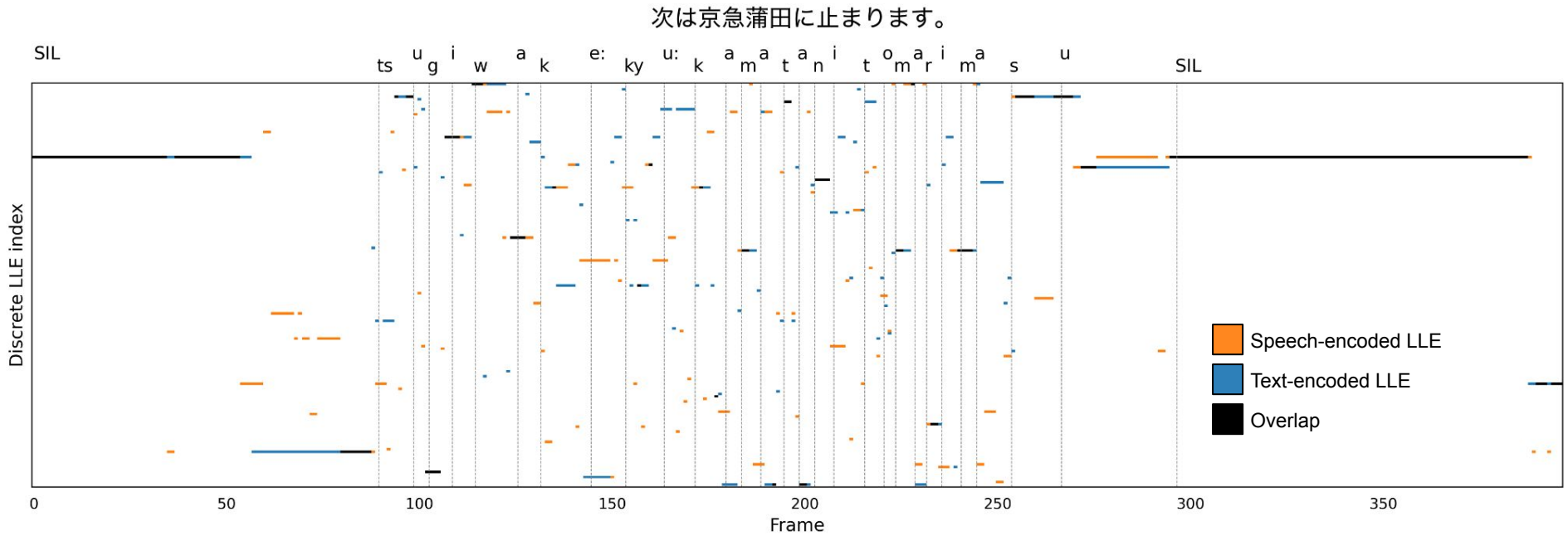
Quality



Similarity

Discrete LLE sequences

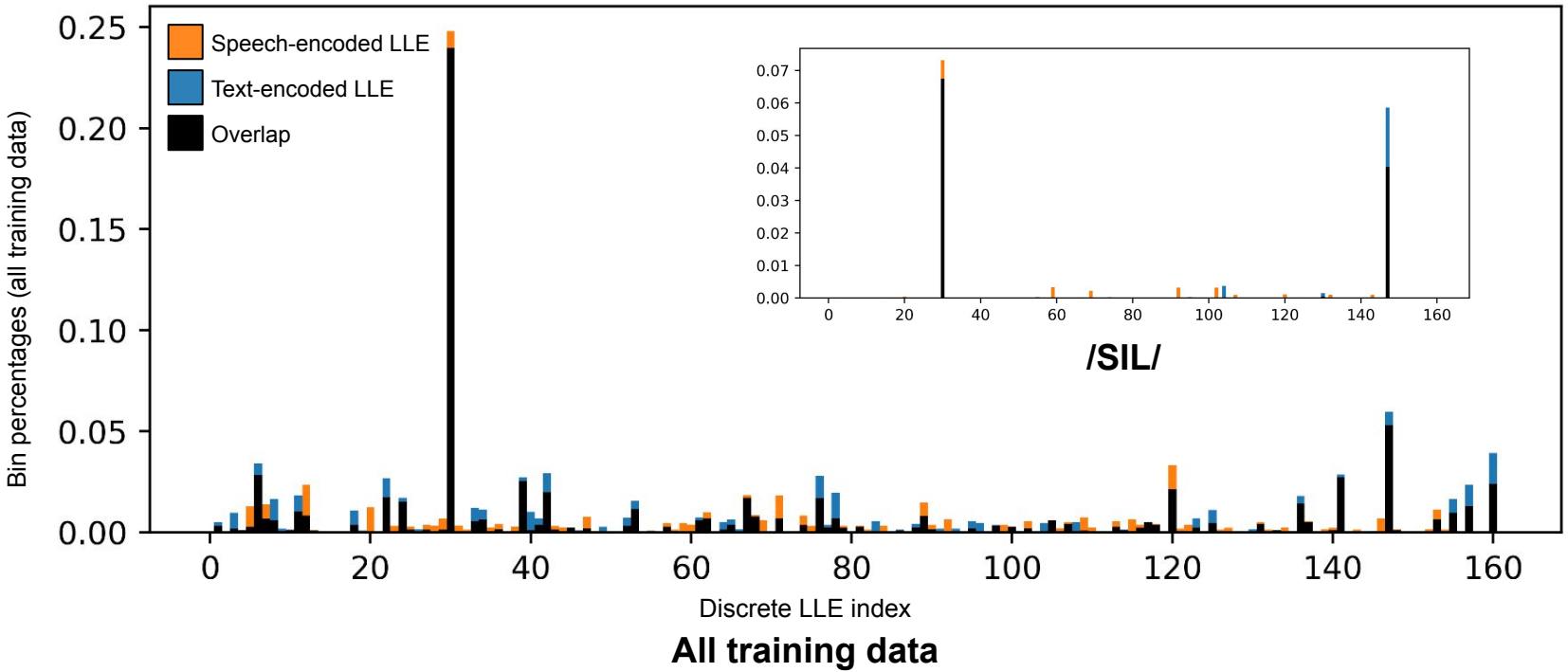
Text Encoder transforms **/SIL/** phoneme to three different discrete vectors. While **Speech Encoder** transforms `**silence'** segments to many more discrete vectors.



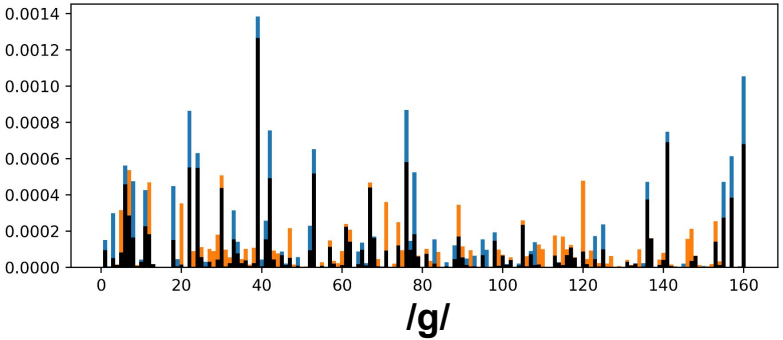
The discrete LLE sequences generated by **speech** and **text encoders** from the respectable **speech** and **text inputs** of the same content overlap by **54.41%** in this particular example utterance.

VQ codes histogram (1)

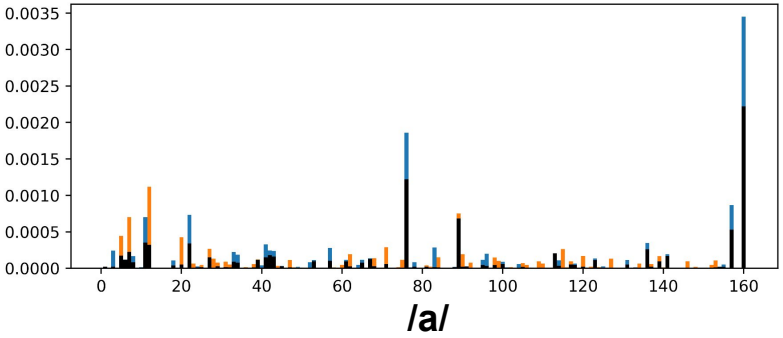
/SIL/ phoneme histogram concentrates on two vectors for both text-encoded and speech-encoded LLE.



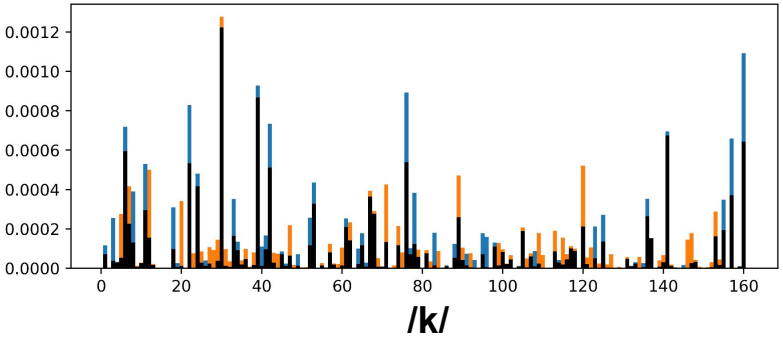
VQ codes histogram (2)



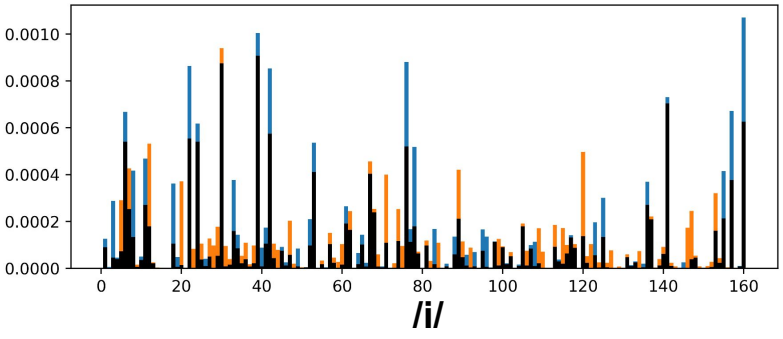
/g/



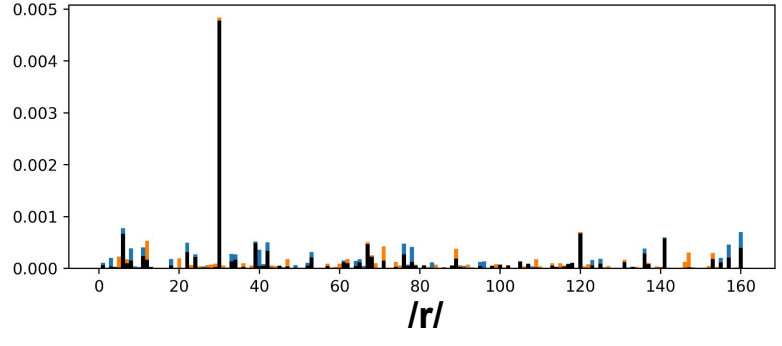
/a/



/k/



/i/



/r/

Conclusion

- **Unified voice cloning TTS/VC** (why?): Convenient and versatile. Moreover by creating a system that operates in two parallel modes (TTS/VC) we created a way to inspect the hidden feature.
  → If we can **observe** it then we can **utilize** it,... somehow (future work)

- **Discrete LLE** (why?): Useful for specific applications (low bit-rate representation, finite discrete vector). Moreover it is easier to analyze the discrete latent space compares with the continuous one.
  → A model with high quality/similarity is required for latent space inspection. As low performance suggests no useful information is learn by LLE.

In summary: VQVAE-based encoders can be used for our unified voice cloning system with acceptable performance and allow more interesting analysis on the linguistic latent space.

Thank you for your attention

PLATINUM SPONSOR
Google

GOLD SPONSORS
科大讯飞 IFLYTEK
SAMSUNG

PARTNERS