Deep learning based voice cloning framework for a unified system of text-to-speech and voice conversion

Luong Hieu-Thi
Main supervisor: Prof. Yamagishi Junichi

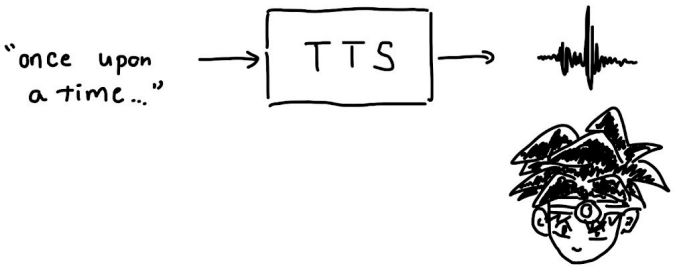
Thesis update

Comparing with the preliminary presentation, the content of the thesis and the presentation are updated to respond to several concerns of examiners:

1) The background chapter was updated with with more content about related prior works of both TTS and VC.
2) More information on research motivations and explanation of the proposed methodologies were added throughout the thesis.
3) General sentence structure, grammar and narrative were refined.

Text-to-speech (TTS)

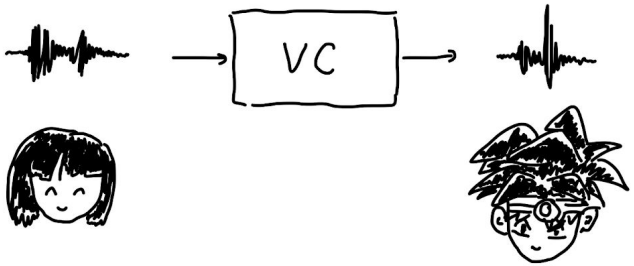
Generating speech with voice of a target speaker from a given text input.



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),...

Motivations

Deep learning based voice cloning framework for a unified system of text-to-speech (TTS) and voice conversion (VC)

Problem 1 - voice cloning framework for TTS and VC

Data efficient method for building TTS or VC system which imitates a target voice. It is a common research theme of both TTS and VC.

Problem 2 - unified system of TTS and VC

Beside for convenience, it is expected to produce speech with highly consistent speaker characteristics between two tasks so they can be used together seamlessly. It is a relatively new research.

Why? TTS is cost effective as text is cheap to produce, but VC can convert content which cannot is difficult to represent by expected written form (e.g. foreign language).

Voice cloning

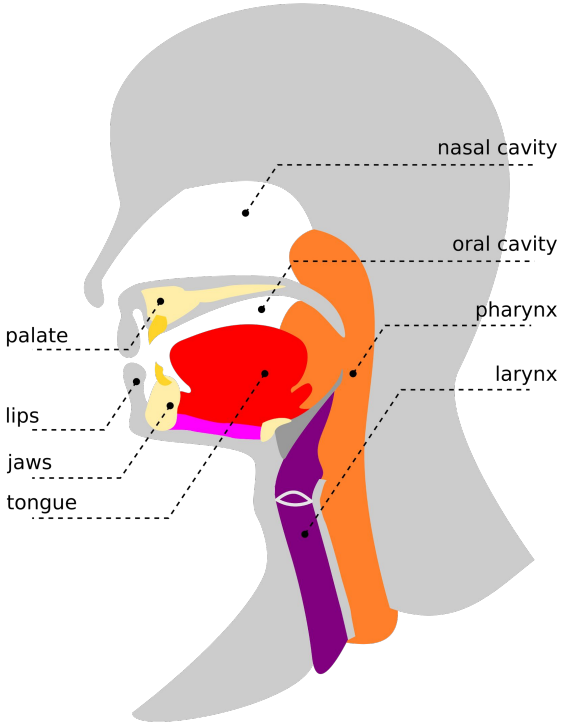
It is not a well-defined term. In pop culture “voice cloning” is loosely used to describe a technology that resembles voice conversion (e.g., James Bond, Detective Conan).

Definition: **any type of speech generation system that imitates voice of a target speaker**. This thesis focuses on voice cloning method for TTS and VC.

Performance evaluations:

- Quality and speaker similarity
- ~~Small footprint and fast computing~~
- Data efficiency:
  - Quantity (small/large amount of data)
  - Quality (transcribed/untranscribed?, clean/noisy?)

…Deep Fake for speech.



TTS voice cloning

How to build a TTS system for a target speaker's voice whose speech data is limited, with or without transcript?

Conventional TTS: system is trained on dozens of hours of transcribed speech data of single speaker → unfit for voice cloning.

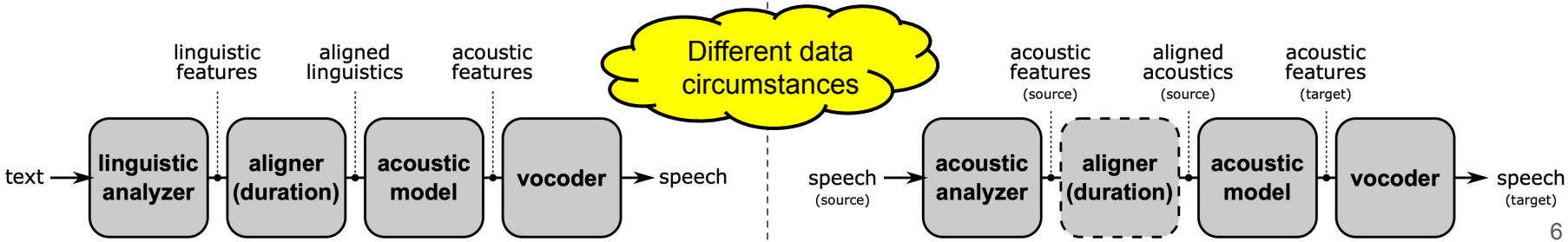
Speaker adaptation: tuning a pre-trained TTS model to transcribed speech (supervised) or untranscribed speech (unsupervised) of a target.

VC voice cloning

How to build a VC system for a target speaker's voice whose data is limited, with or without parallel utterances of source speaker?

Parallel VC: using parallel utterances of source and target speakers (duration mismatch) to train the VC model → expensive, limited data.

Non-parallel VC: using non-parallel utterances to train the VC model → convenient, less demanding, more usable data.



A versatile/unified TTS/VC system

A versatile TTS/VC system

A system which is capable of cloning voices of target speakers whose data are under varied circumstances (e.g., transcribed speech, untranscribed speech, parallel, non-parallel, small amount of data, a large amount of data).

A unified TTS/VC system

One system which can act as both TTS and VC with highly consistent performance with a target voice so they could be used together for a single task (e.g., video games).

Can we create a versatile/unified state-of-the-art (SOTA) TTS/VC system?

Application of a unified TTS/VC system

e.g,, video game dialogs

Using TTS to generate majority of speech that conveys contents using simple expression (text is easier to modify).

e.g., video game cutscenes

Using VC to generate speech that conveys unconventional expression or difficult to represent in expected written forms.

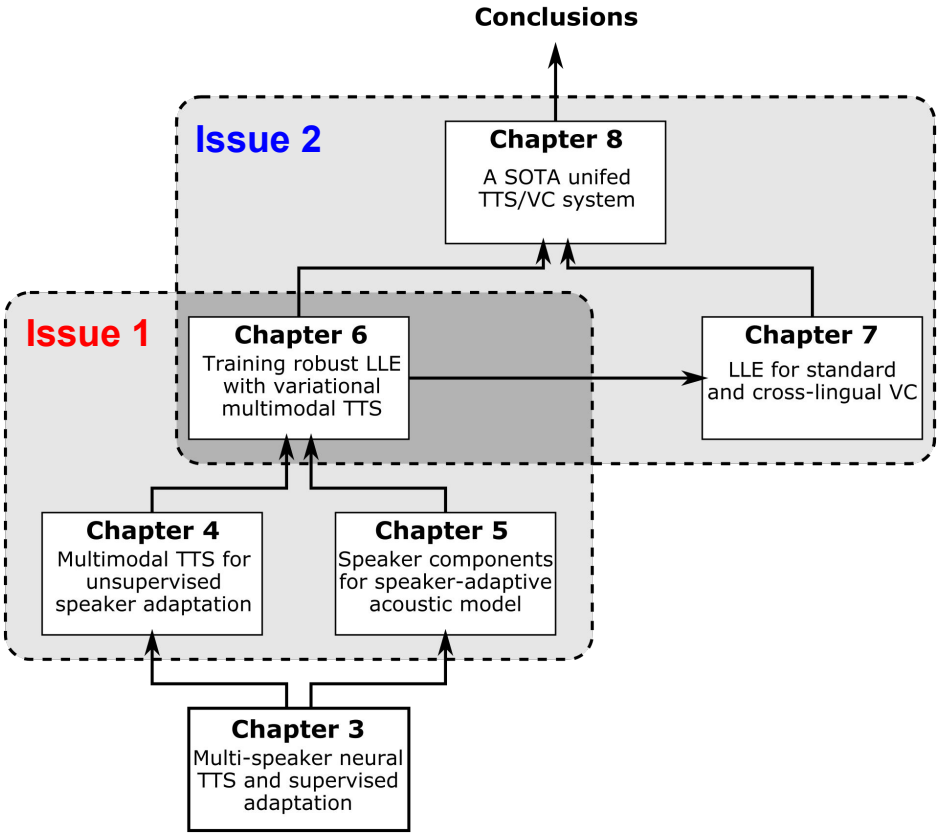
A unified TTS/VC system is expected to have consistent performance between the two, so we can switch between them seamlessly.

Thesis outline

This thesis tackles two major issues. Each issue spreads over several chapters.

Issue 1: developing a versatile speaker adaptation method for neural TTS which can work with both transcribed and untranscribed speech

Issue 2: developing a unified voice cloning system of TTS and VC with high quality and data efficiency. More importantly, high speaker consistency between two so that we can use them together seamlessly.



Premise

Multi-speaker neural TTS and supervised adaptation

Chapter 2 and 3 of the thesis

Neural TTS model

Neural network: an input goes through a series of non-linear transformations to become the desired output.

Using gradient descent and backpropagation to optimize network parameters.

Neural TTS model: the input **x** is a text representation and the output is a speech representation **y**

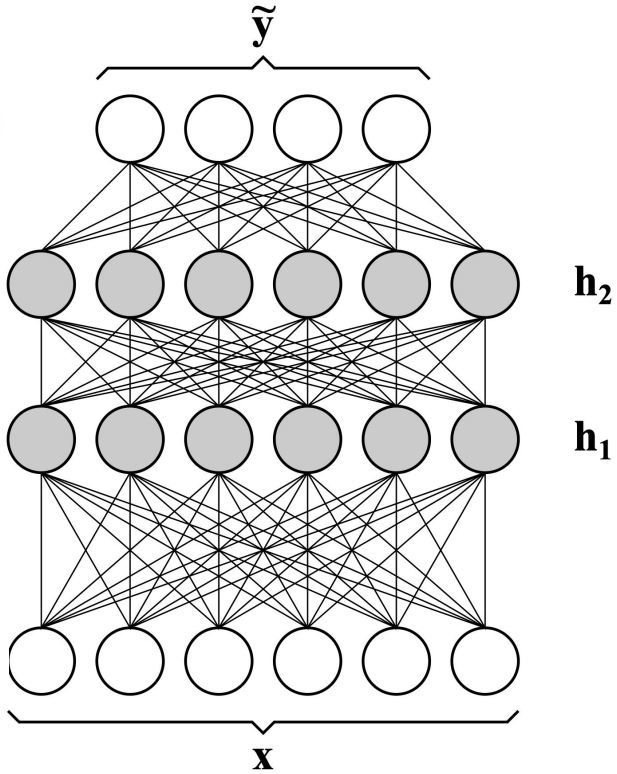
$$\tilde{\boldsymbol{y}} = TTS(\boldsymbol{x}; \boldsymbol{\Theta}^{tts})$$

Need a large amount of transcribed speech data (e.g., 10 hours) from a single speaker.

$$\tilde{\boldsymbol{y}} = f(\boldsymbol{W}_L \boldsymbol{h}_{L-1} + \boldsymbol{c}_L)$$

$$\boldsymbol{h}_l = f(\boldsymbol{W}_l \boldsymbol{h}_{l-1} + \boldsymbol{c}_l)$$

$$\boldsymbol{h}_1 = f(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{c}_1)$$



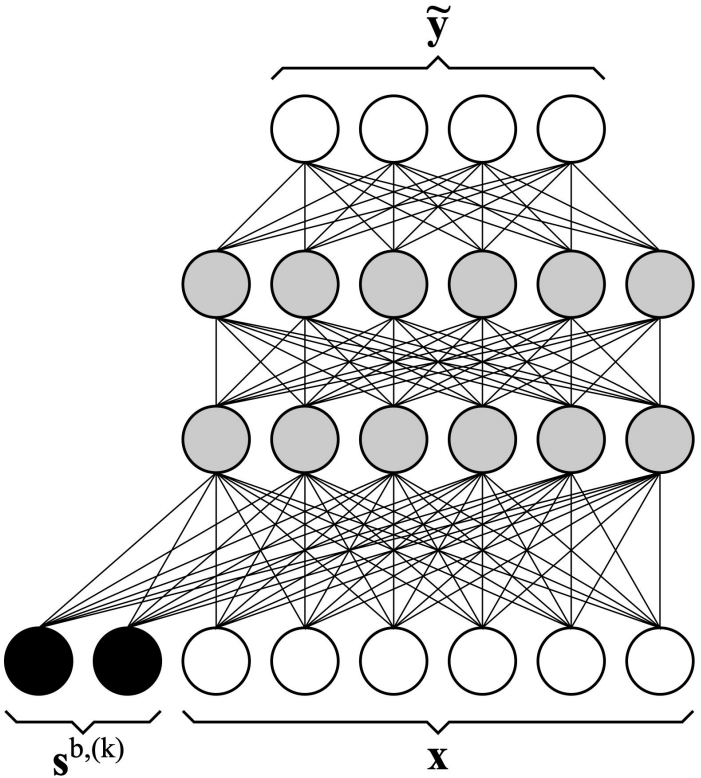
Multi-speaker acoustic model (1)

Motivation: combined small amount of transcribed speech (e.g. 10 minutes) of multiple speakers to train a stable TTS model for all.

Multi-speaker acoustic model: augmenting a speaker embedding into the linguistic input. Within a single utterance, linguistic input **x** changes every frame, but speaker code $\mathbf{s}^{b,(k)}$ stays the same.

Different types of speaker embedding: one-hot vector, random vector, discriminant code, i-vector, x-vector,...

See **Chapter 3** of the thesis for more details.

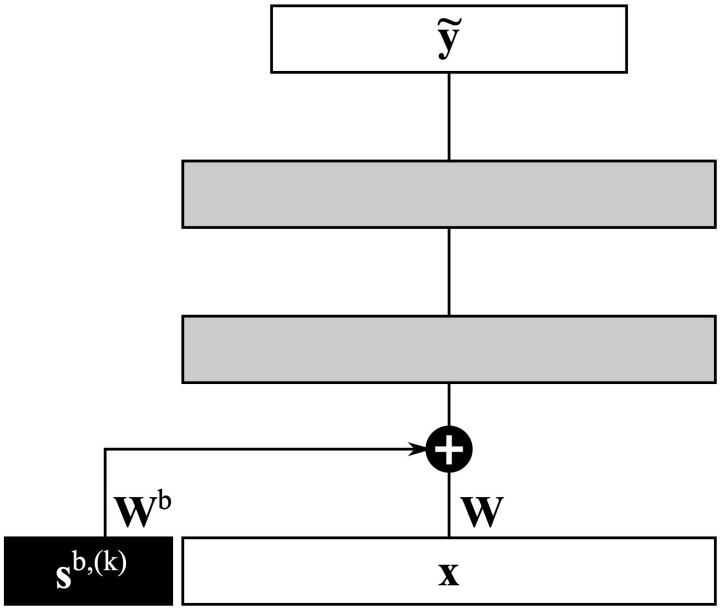


Multi-speaker acoustic model (2)

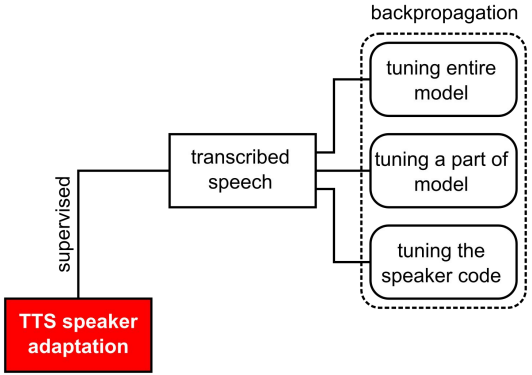
Multi-speaker acoustic model mechanism: the speaker code changes the layer bias which changes the function modeled by network:

$$\boldsymbol{h}_1 = \tanh(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{c}_1 + \boldsymbol{W}^b\boldsymbol{s}^{b,(k)})$$
$$\boldsymbol{h}_1 = \tanh(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{c}_1 + \boldsymbol{b}^{(k)})$$
$$\boldsymbol{h}_1 = \tanh(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{c}_1^{(k)})$$

We use $\mathbf{s}^{b,(k)}$ that jointly trained with the acoustic model, and refer to it as **speaker bias code**



Supervised speaker adaptation



Motivation: adapt a pretrained TTS model to voice of an unseen speaker whose data consists of a small amount of transcribed speech → similar to multi-speaker task but faster.

Adaptation with backpropagation: we can tune a part of or the entire model.

Tuning more parameters → better speaker similarity, more vulnerable to overfitting.

Tuning less parameters → worse speaker similarity, less vulnerable to overfitting.

⇒ Strike for balance

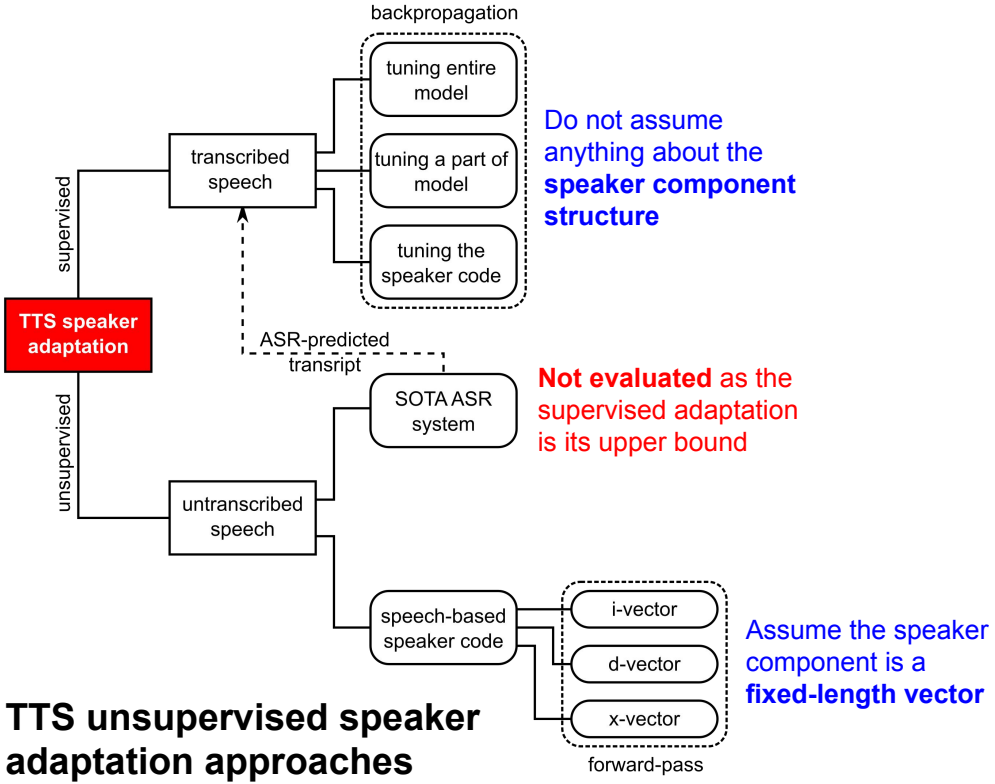
TTS supervised speaker adaptation approaches

Part 1

Versatile speaker adaptation method for TTS

Chapter 4, 5 and 6 of the thesis

Unsupervised speaker adaptation



Do not assume anything about the **speaker component structure**

Not evaluated as the supervised adaptation is its upper bound

Assume the speaker component is a **fixed-length vector**

TTS unsupervised speaker adaptation approaches

Motivation: adapt a pretrained TTS model to voice of an unseen speaker whose data consists of a small amount of untranscribed speech → significantly reduce cost of creating new voice for the TTS model.

Inoue, Katsuki, et al. "Semi-Supervised Speaker Adaptation for End-to-End Speech Synthesis with Pretrained Models." Proc. *ICASSP*, 2020.

Cooper, Erica, et al. "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings." Proc. *ICASSP*, 2020.

Related work: d-vector for unsupervised adaptation

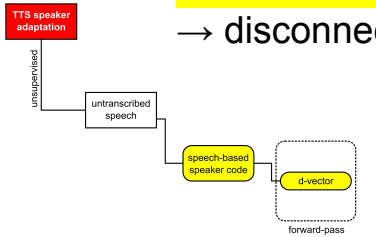
Extracted speaker embedding from several speech samples of a target speaker to change the voice of the generated speech.

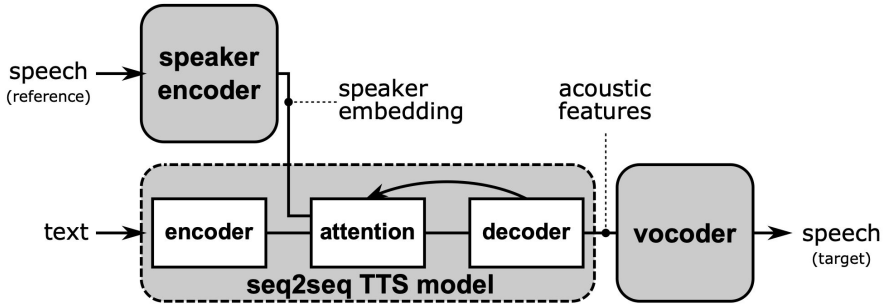
Pros:

- Fast adaptation (forward-pass)
- Reliable quality (overfitting resistance)

Cons:

- Low speaker similarity, low scalability.
- **Speaker encoder** is trained separately from the rest of model.
  → disconnect pipeline





Jia, Ye, et al. "Transfer learning from speaker verification to multispeaker text-to-speech synthesis." *Advances in neural information processing systems*. 2018.

Multimodal neural TTS

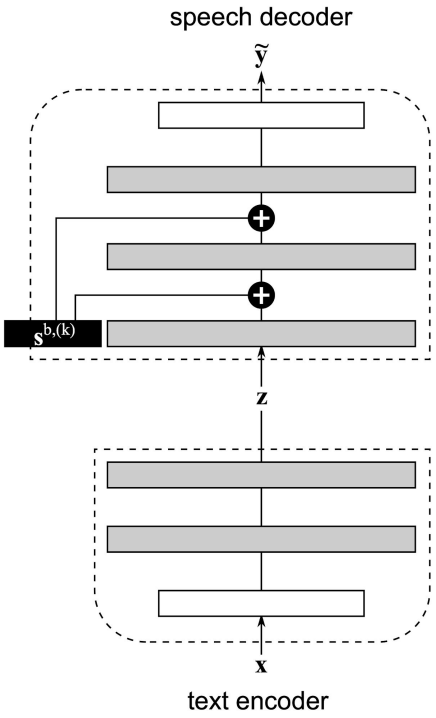
We want a backpropagation-based unsupervised adaptation method, without having to train an SOTA ASR system.

Proposal: train an auxiliary speech encoder along with a conventional TTS model so we could use it to adapt with untranscribed speech.
First, split the TTS model into **text encoder** and **speech decoder**. Then add **speech encoder.**

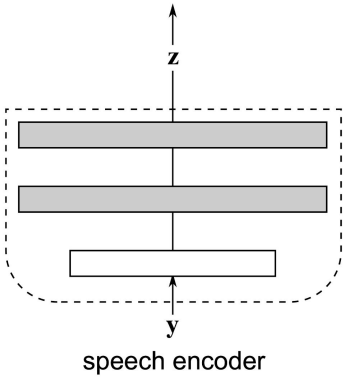
$$\boldsymbol{z}^T = TEnc(\boldsymbol{x}; \phi^T)$$
$$\tilde{\boldsymbol{y}}^T = SDec(\boldsymbol{z}^T; \theta^{\mathrm{S/core}}, \theta^{\mathrm{S/spk},(k)})$$

Typically, we train it with TTS loss:

$$\mathrm{loss}_{tts} = Cost(\tilde{\boldsymbol{y}}^T, \boldsymbol{y})$$



Latent feature **z** is called **Latent Linguistic Embedding (LLE)**



Crossmodal adaptation

Add a **speech encoder** which acts as a substitute for **text encoder**:

$$\boldsymbol{z}^S = SEnc(\boldsymbol{y}; \phi^S)$$

$$\tilde{\boldsymbol{y}}^S = SDec(\boldsymbol{z}^S; \theta^{\mathrm{S/core}}, \theta^{\mathrm{S/spk},(k)})$$

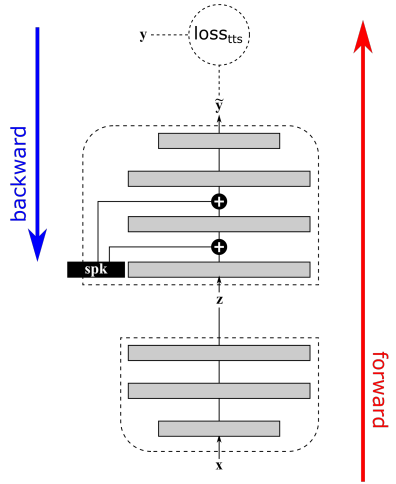
Typically we can train it with STS loss:

$$\mathrm{loss}_{sts} = Cost(\tilde{\boldsymbol{y}}^S, \boldsymbol{y})$$

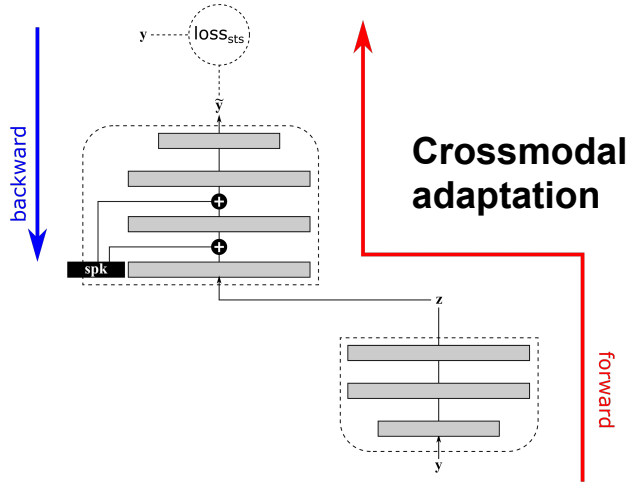
However this does not guarantee a consistent latent space between text and speech encoders.

But if we can train such multimodal network, we can use STS stack to estimate a new $\mathbf{s}^{\mathrm{b},(\mathrm{r})}$ for the r-th target speaker with untranscribed speech:

$$\mathrm{loss}_{adapt} = \mathrm{loss}_{sts}$$



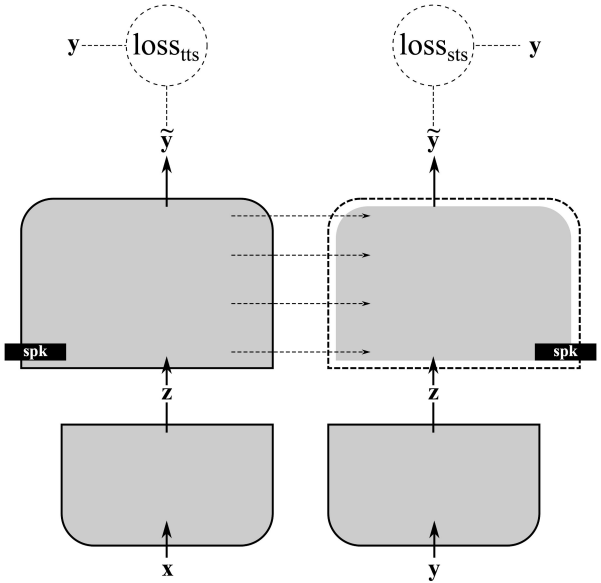
Conventional adaptation



Crossmodal adaptation

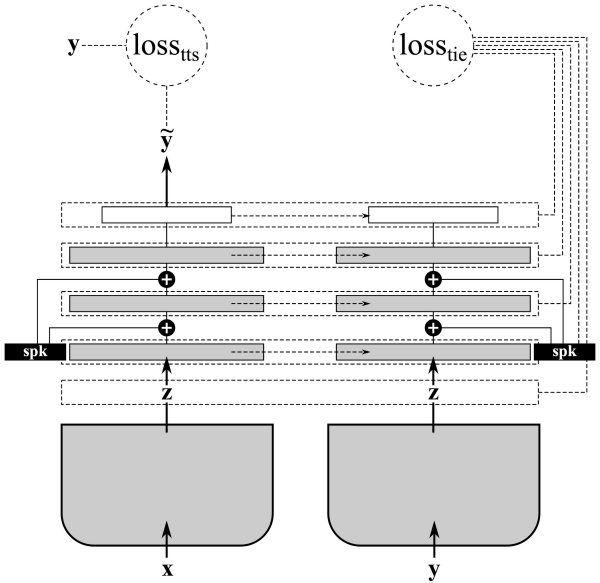
Multimodal learning methods

$$\text{loss}_{tie} = \sum_{\substack{l=1 \\ l \in \boldsymbol{F}}}^{L} \text{distance}(\boldsymbol{h}_l^{tts}, \boldsymbol{h}_l^{sts})$$



Joint-goal $\text{loss}_{train} = \text{loss}_{tts} + \alpha\ \text{loss}_{sts}$

The outputs of each stack are jointly optimized toward the same goal.

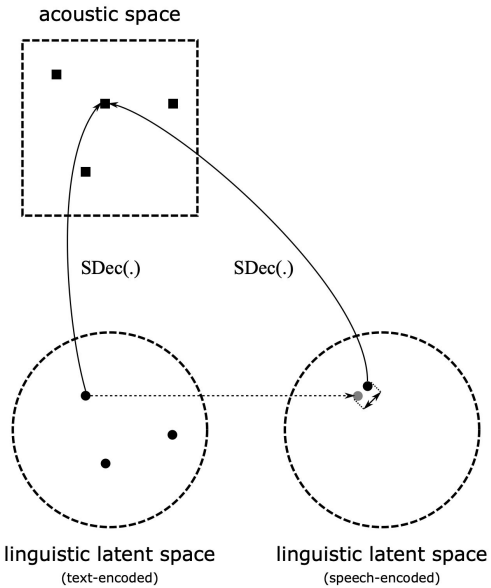


Tied-layer $\text{loss}_{train} = \text{loss}_{tts} + \beta\ \text{loss}_{tie}$

The hidden layers spaces of each stack are optimized toward each other.

There are also **step-by-step** and **stochastic training**, see **Chapter 4** of the thesis for more details.

Multimodal network & AE



The **acoustic decoder** transforms a sample in the shared linguistic latent space to a sample in acoustic space.

What do we want? Train the speech-encoded latent space that is **identical** to text-encoded latent space → difficult/impossible

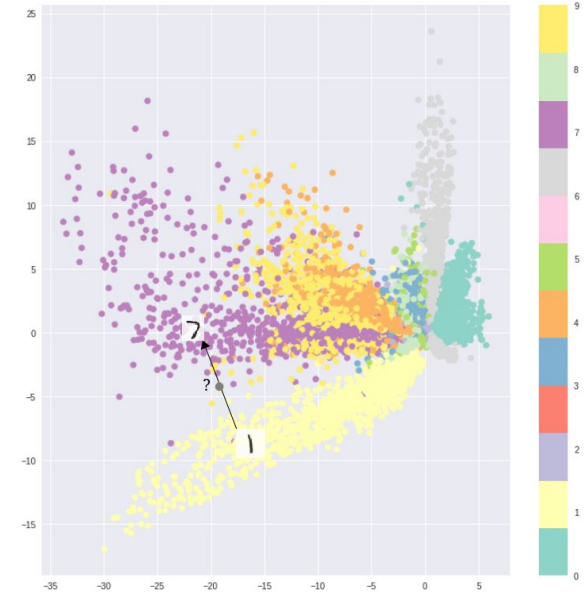
What can we do? Train the speech-encoded latent space to **approximate** the text-encoded latent space → feasible

What does that mean? Points close in shared latent space produces point close in acoustic space → we need continuity

What is the problem? Discrete latent point + data scarcity = sparse/discontinuous latent space (autoencoder problem)

→ not robust to unseen samples.

AE & VAE



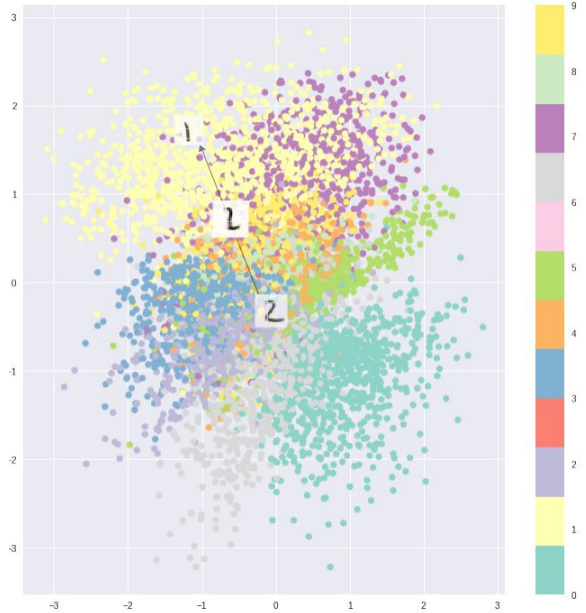
Latent space of an autoencoder trained on MNIST dataset which has discontinuity and incompleteness.

Image credit:
https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Autoencoder latent space is continuous for the most part, but for region in which we do not have any training samples, we are unsure about their behavior.

Variational autoencoder latent space, in the other hand, has better continuous and more completed thanks to its structure.

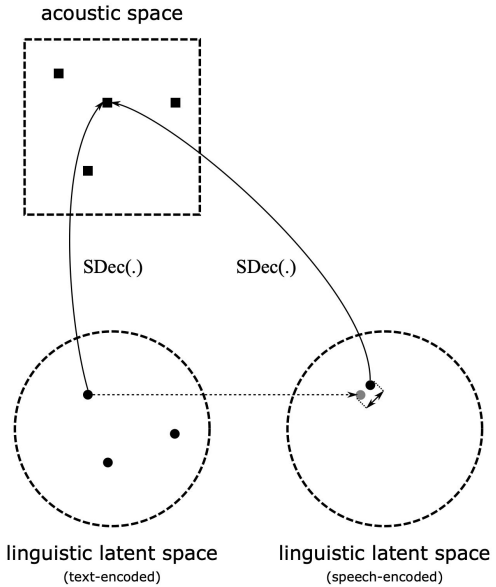
Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." *arXiv preprint arXiv:1312.6114*, 2013.



Latent space of a variational autoencoder trained on MNIST dataset which has better continuity and completeness than vanilla autoencoder.

Image credit:
https://towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

Multimodal network & VAE



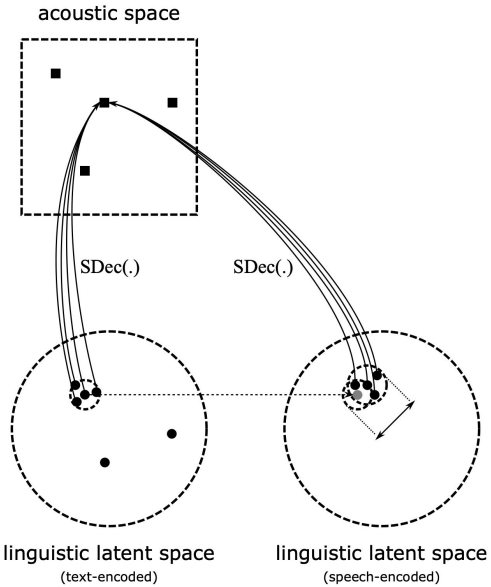
The shared linguistic latent space with AE-like structure

Given the previous observation we use a VAE-like structure to improve the continuity and consistency of the shared linguistic latent space.

How does VAE-like structure help with the continuity? sampling process acts as artificial data generation which complied with continuity policy.

How does VAE-like structure help the speech-encoded latent space approximate the text-encoded latent space? It enables the use of density-wise distortion function instead of pointwise function*.

* We can **assume** simple distribution with AE-like as well, but it is unstable to do so. VAE-like **forces** the latent space to take simple distribution form.



The shared linguistic latent space with VAE-like structure

Variational multimodal neural TTS

To increase the continuity of linguistic latent space, we change the output of encoders into distributions instead of points. Text-to-speech stack:

$$\boldsymbol{z}^T \sim TEnc(\boldsymbol{x}; \phi^T) = p(\boldsymbol{z}|\boldsymbol{x})$$

$$\tilde{\boldsymbol{y}}^T = SDec(\boldsymbol{z}^T; \theta^{\mathrm{S/core}}, \theta^{\mathrm{S/spk},(k)})$$

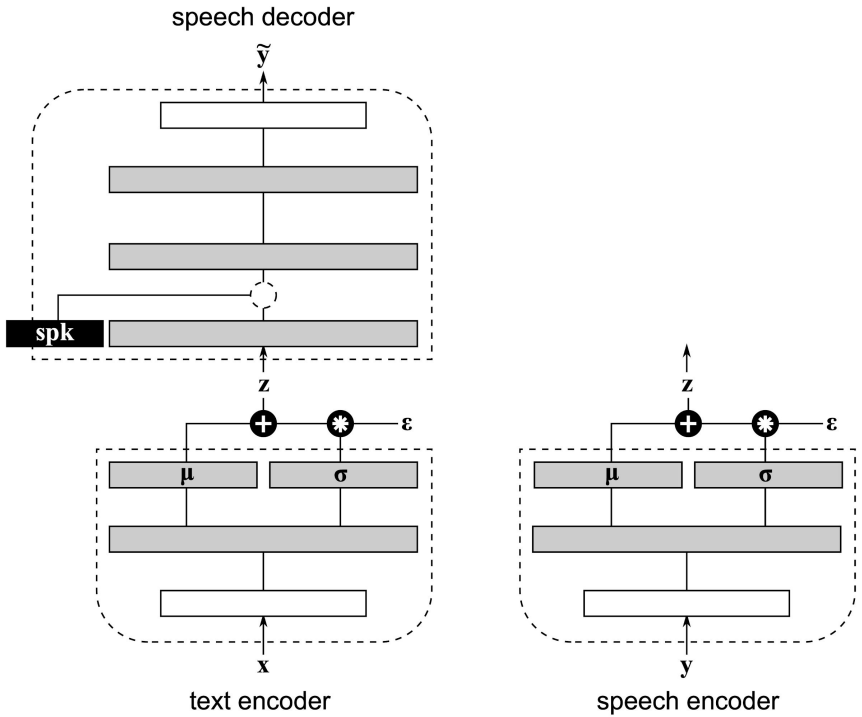
Speech-to-speech stack:

$$\boldsymbol{z}^S \sim SEnc(\boldsymbol{y}; \phi^S) = q(\boldsymbol{z}|\boldsymbol{y})$$

$$\tilde{\boldsymbol{y}}^S = SDec(\boldsymbol{z}^S; \theta^{\mathrm{S/core}}, \theta^{\mathrm{S/spk},(k)})$$

To train this network, we make encoders output mean and std of an isotropic Gaussian then applied reparameterization trick (VAE-inspired)

$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1)$$



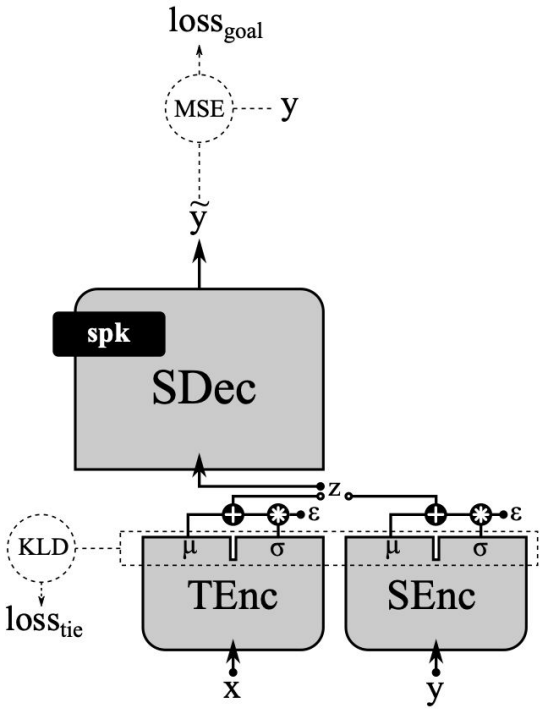
Training a variational multimodal neural TTS

The variational multimodal neural TTS can be trained using the same method as standard one with joint-goal and tied-layer learning.

Given a VAE-like structure we could use a special tied-layer setup for training using Kullback-Leibler divergence:

$$\text{loss}_{train} = \text{loss}_{tts} + \beta\ \text{loss}_{tie}$$

$$\text{loss}_{tie} = L_{KLD}(TEnc(\boldsymbol{x}), SEnc(\boldsymbol{y}))$$

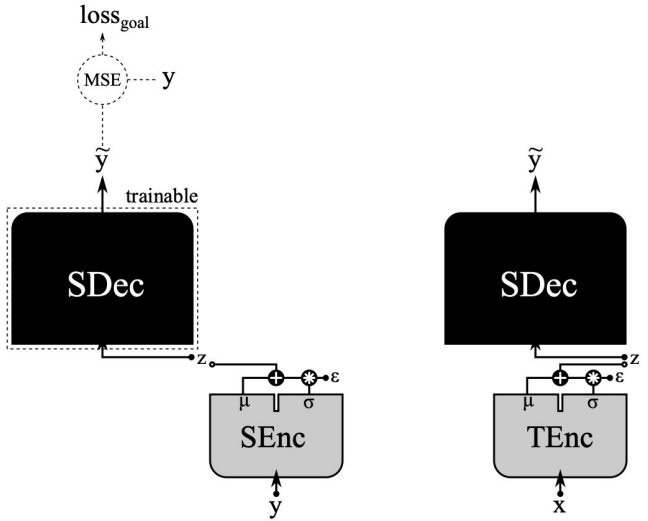


Adapting by fine-tuning entire module

Fine-tuning the speaker embedding is too restrictive but it is necessary to prevent overfitting.

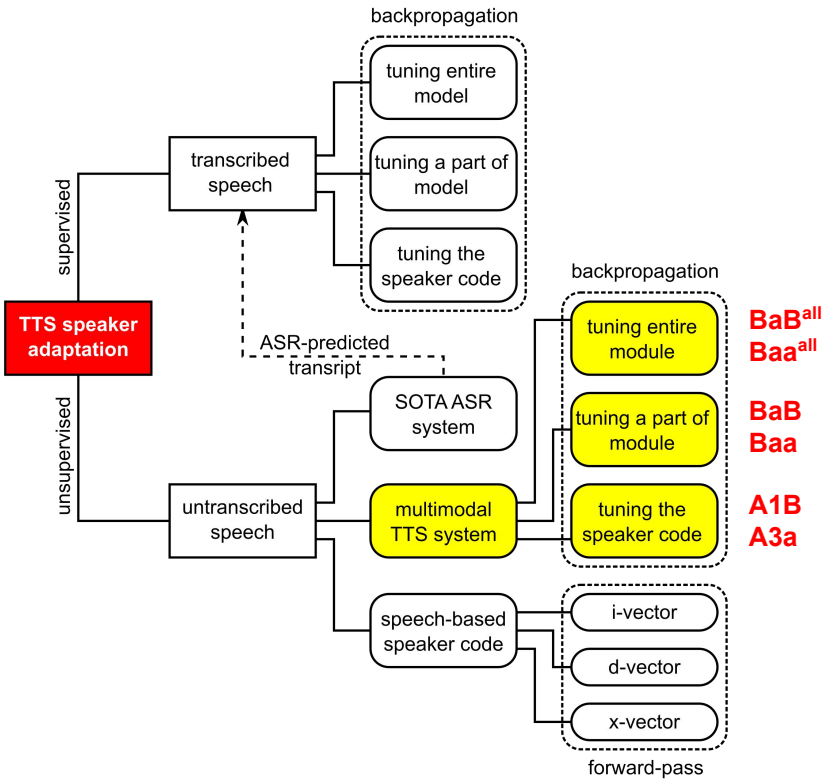
Hypothesis: If the speech-encoded latent space is a good approximation of the text-encoded latent space, we can perform crossmodal adaptation by fine-tuning the entire module without worrying about overfitting:

$$\text{loss}_{adapt} = \text{loss}_{sts}$$



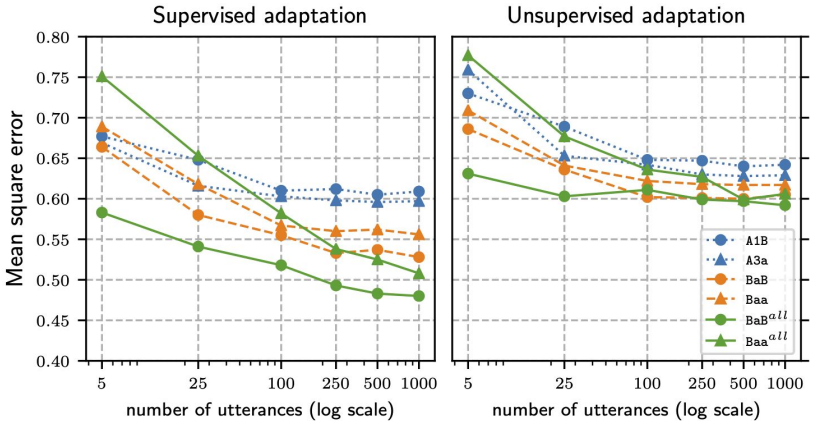
Adaptation and Inference stages

EXP: speaker adaptation with multimodal TTS - Systems

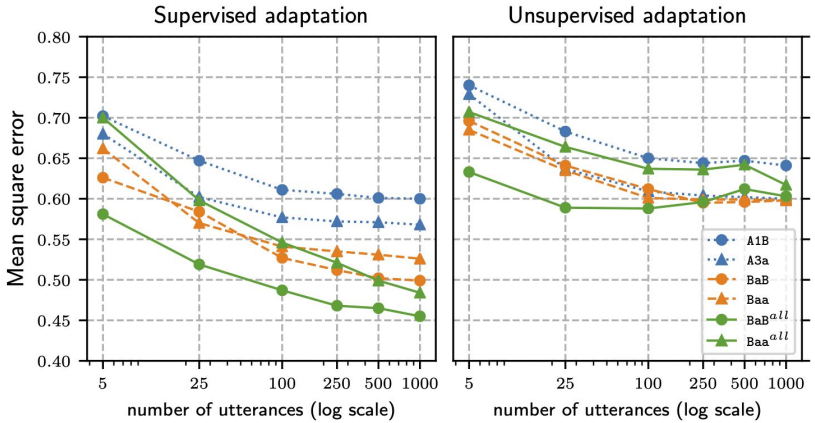


Strategy	Speaker layer	Speaker code		Adaptable layer(s)
		Scaling	Bias	
A1B	A1	-	full	At one layer
A3a	A3	128	128	At one layer
BaB	B[1-8]	-	full	At multiple layers
Baa	B[1-8]	64	64	At multiple layers
BaB^{all}	B[1-8]	-	full	Speech decoder
Baa^{all}	B[1-8]	64	64	Speech decoder

EXP: speaker adaptation with multimodal TTS - Objective



(a) Female speaker



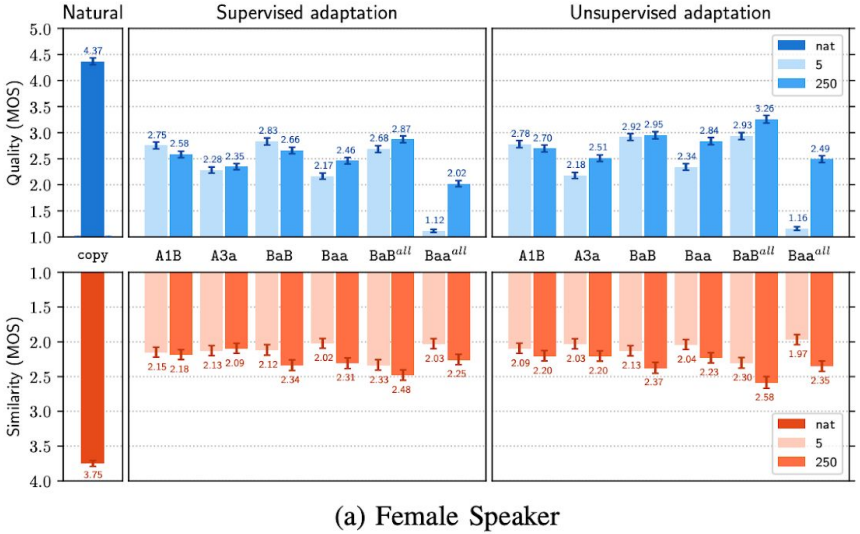
(b) Male speaker

BaB^{all} has the best performance in all data points, even in 5-utterance case.

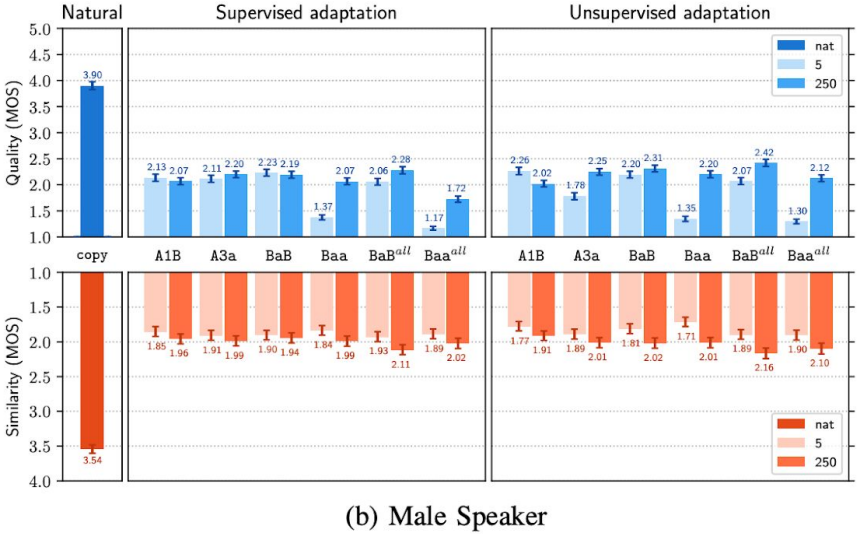
Baa^{all} has the worst performance in 5-utterance case, but improved significantly in 1000-utterance case.

→ Given a good pretrained model, adaptation does not become overfitting even when data is limited.

EXP: speaker adaptation with multimodal TTS - Subjective



(a) Female Speaker



(b) Male Speaker

BaB^{all} is the best in general, with its unsupervised strategy has better subjective evaluation than its supervised. Even though objective results show the opposite.

→ It is tricky to evaluate perception task with objective evaluation.

Review: Part 1

Contribution

Successfully proposed a novel versatile/high-scalability speaker adaptation method for TTS.

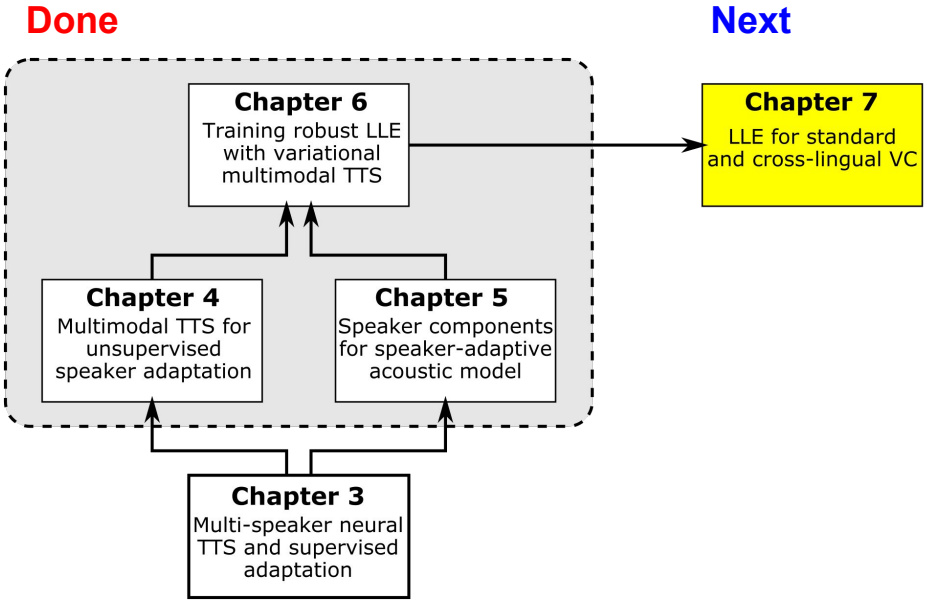
→ adapting with transcribed or untranscribed speech, resistant to overfitting when amount of data is small, but quality improves significantly when more data becomes available.

Remaining challenges

It is a proof-of-concept (POC) so the general performance is not very high.

What's next

We move on to develop a unified voice cloning system for TTS and VC



Part 2

Bootstrapping VC from TTS

Chapter 7 of the thesis

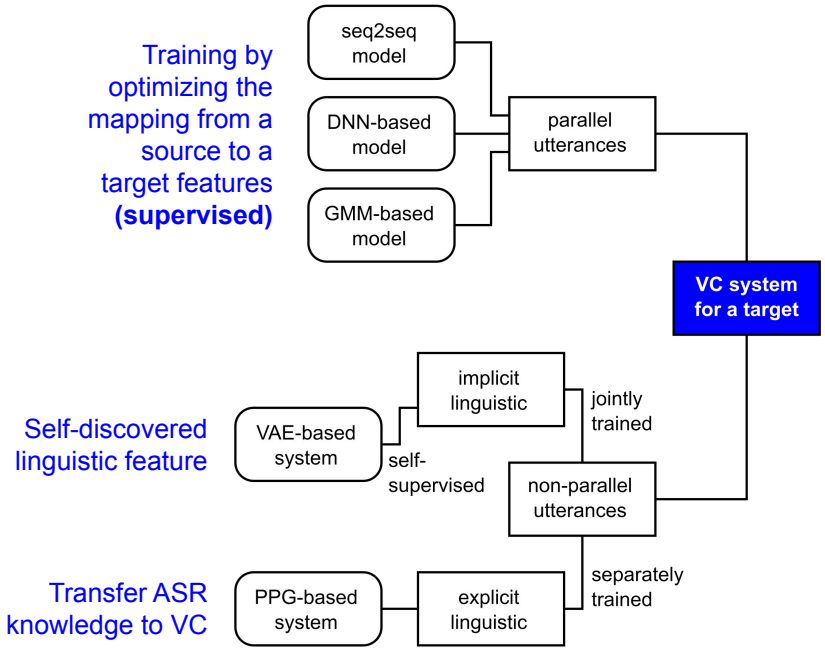
Complementary of TTS and VC

TTS and VC can be seen as different interfaces for generating speech which are useful for different application scenarios.

If we can create TTS and VC systems with consistent performance, they can be used together for the same task → even more useful (synergy)

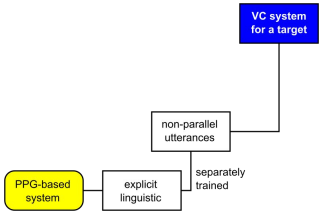
If we create TTS and VC systems using the same methodology and setup, it is more likely to have similar performance.

→ bootstrapping VC system from TTS



Existing VC approaches

Related work: Phonetic posteriorgram (PPG) for VC



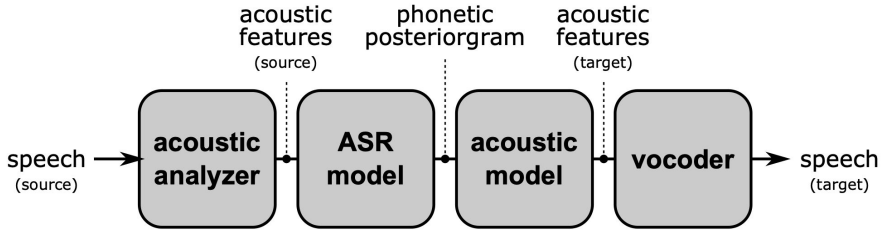
Using ASR to extract PPG, then using PPG to train a TTS-like acoustic model → transfer knowledge learned by ASR to VC.

Pros:

- high quality and speaker similarity when data of target speaker is sufficient.
- Non-parallel/any-to-one VC.

Cons:

- required state-of-the-art (SOTA) ASR system trained on large-scale transcribed corpus.
- The ASR model and the speech generation acoustic model are trained separately.
  → disconnected pipeline.



Sun, Lifa, et al. "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training." *2016 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2016.

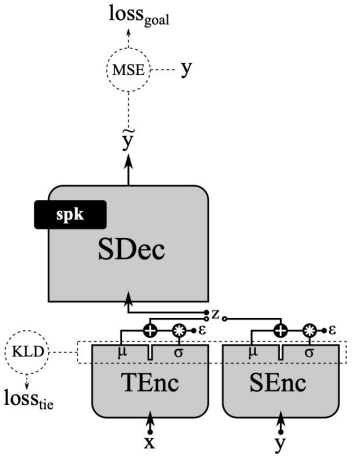
Latent linguistic embedding for VC

The variational multimodal neural TTS trained previously might be used for voice conversion as well.

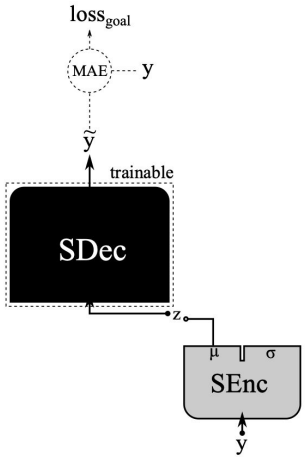
Hypothesis: if the latent linguistic embedding is speaker-disentangled and consistent enough between text and speech encoders, then we can used speech-to-speech stack as VC

$$\text{loss}_{adapt} = \text{loss}_{sts}$$

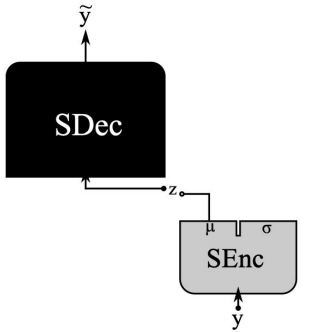
TACTIC using mean-value LLE for adaptation to focus on learning fine-grained details instead of generalization.



(1) Pretraining



(2) Adaptation



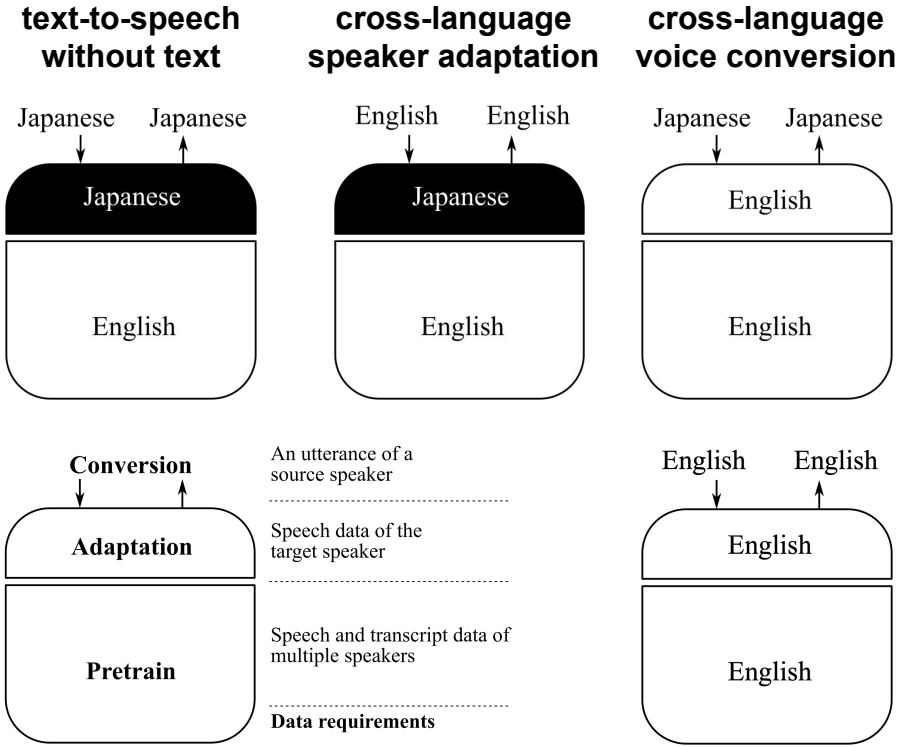
(3) Converting

Cross-lingual VC scenarios

The VC system is built in three stages:

1) **Training** the multimodal neural TTS using transcribed multi-speaker corpus
2) **Adapting** to target speaker using untranscribed speech
3) **Converting** an utterance of arbitrary source speaker

The imbalance in data demand at each stage can be utilized for cross-lingual voice conversion.

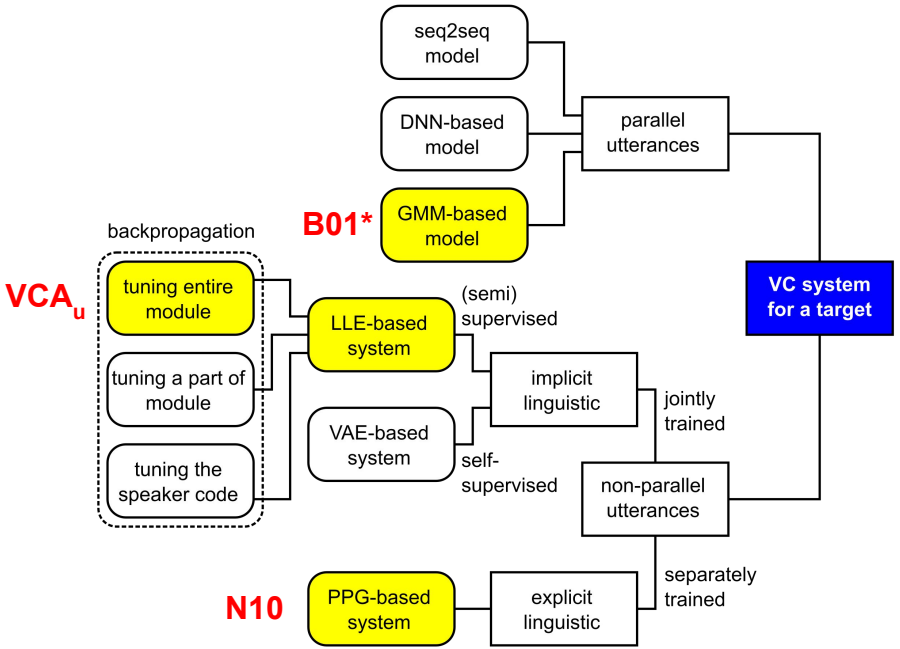


EXP: LLE for intra-language VC - Systems

Re-enactment of SPOKE task of Voice Conversion Challenge 2018 (VCC2018) which built VC system for 4 speakers with 81 utterances from each (approx. 5 minutes)

Systems:

- **B01**: baseline system, 3rd in quality and 6th in similarity
- **N10**: **the best system** in both measurements. A realization of PPG-based VC
- **VCA$_u$**: the proposed any-to-one VC system using LLE.

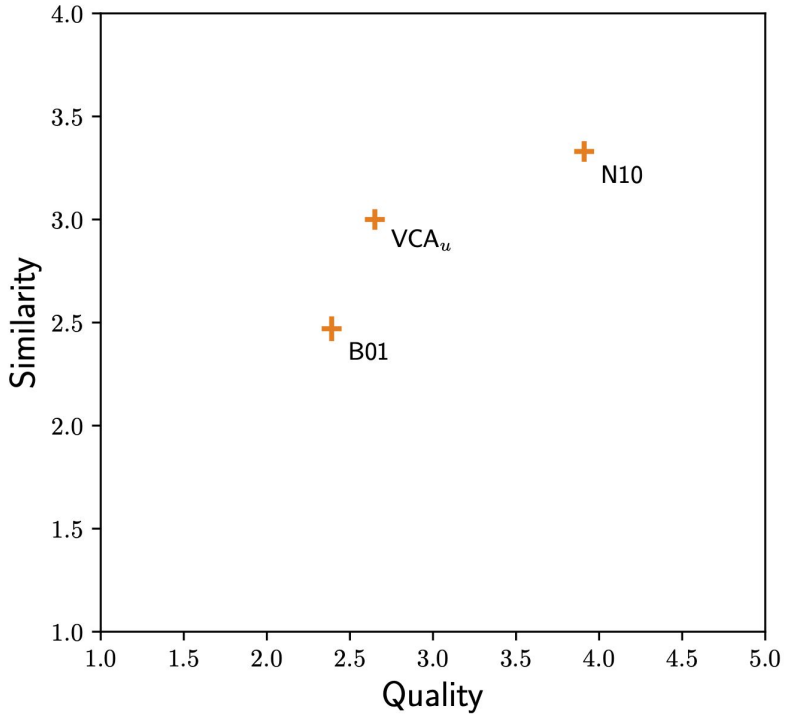


EXP: LLE for intra-language VC - Subjective

Evaluation:

The proposed system **VCA$_u$** has high speaker similarity and quality than **B01**, but is not as good as the best system **N10** yet.

→ validating the proposed bootstrapping VC from TTS framework.



Subjective evaluation

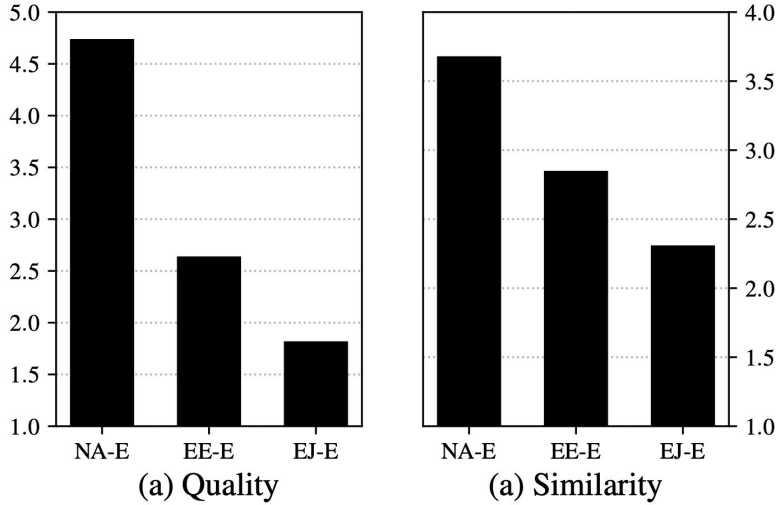
Lines indicate 95% confidence interval

EXP: LLE for cross-language VC

Target speakers are two bilingual (Japanese/English) speakers. The model is adapted with 400 utterances per speaker per language.

EJ-E < EE-E as expected → cross-language speaker adaptation is working to some extent.

Experiments on cross-language VC (**EE-J** and **EJ-J**) received similar results, which provide the POC for using LLE-based VC in those cross-lingual scenarios.



Subjective evaluation

See **Chapter 7** of the thesis for more details.

Review: Part 2

Contribution

Successfully proposed a framework to transfer TTS system to train a flexible VC system.

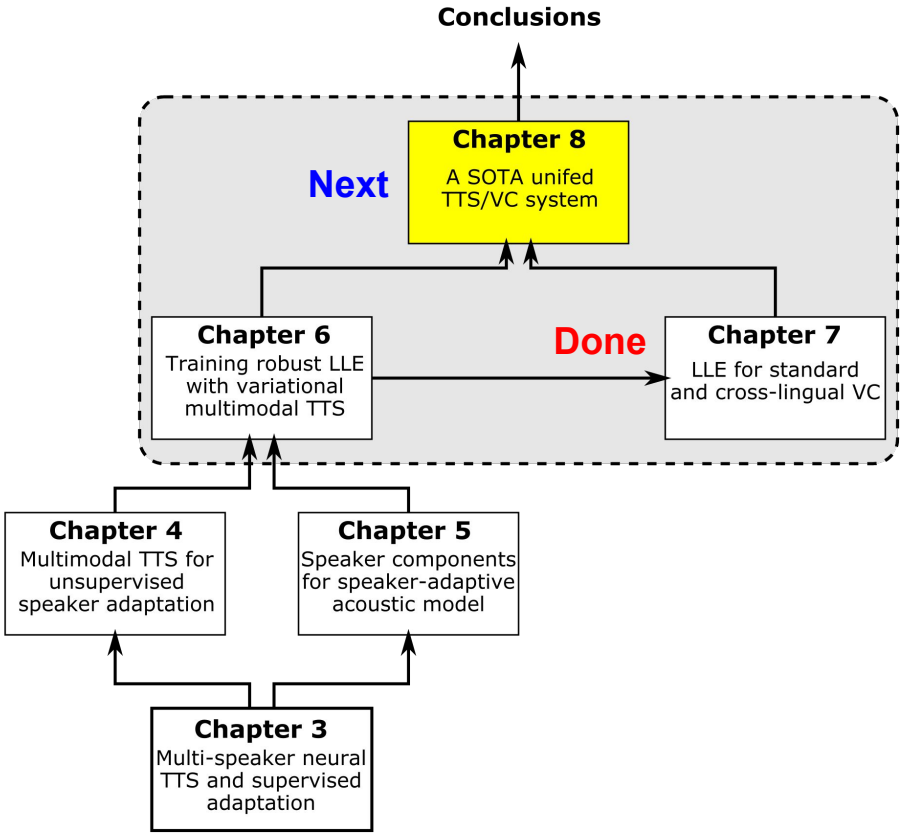
→ any-to-one/cross-lingual VC system, improve quality of converted speech, unite TTS and VC methodology and functional modules.

Remaining challenges

Quality is not high enough (POC), TTS and VC are still handled separately.

What's next

A unified/high-performance voice cloning system of TTS and VC



Related work: Voice transformer network

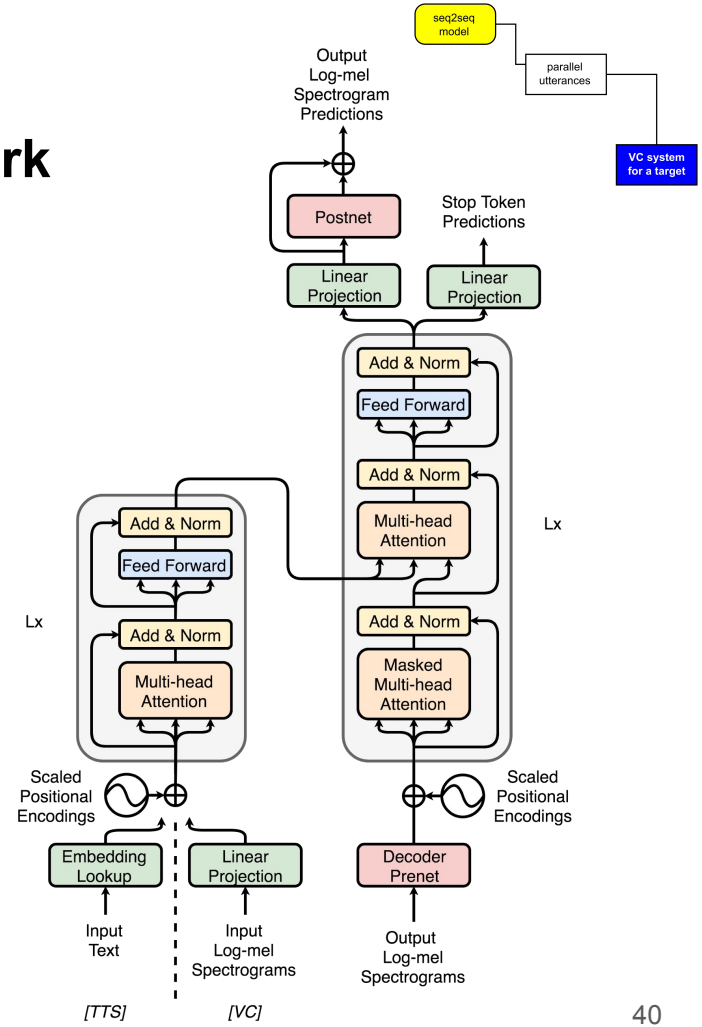
Bootstrapping a VC transformer system from a TTS transformer system.

Similarities:

- (unspecific) multimodal structure with text encoder, speech encoder and speech decoder.
- Pretrain the model with TTS corpus.

Differences:

- Step-by-step training instead of jointly training.
- Supervised speaker adaptation (parallel utterances) instead of unsupervised adaptation (non-parallel utterance).
- E2E/Transformer architecture



Huang, Wen-Chin, et al. "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining." *arXiv preprint arXiv:1912.06813* (2019).

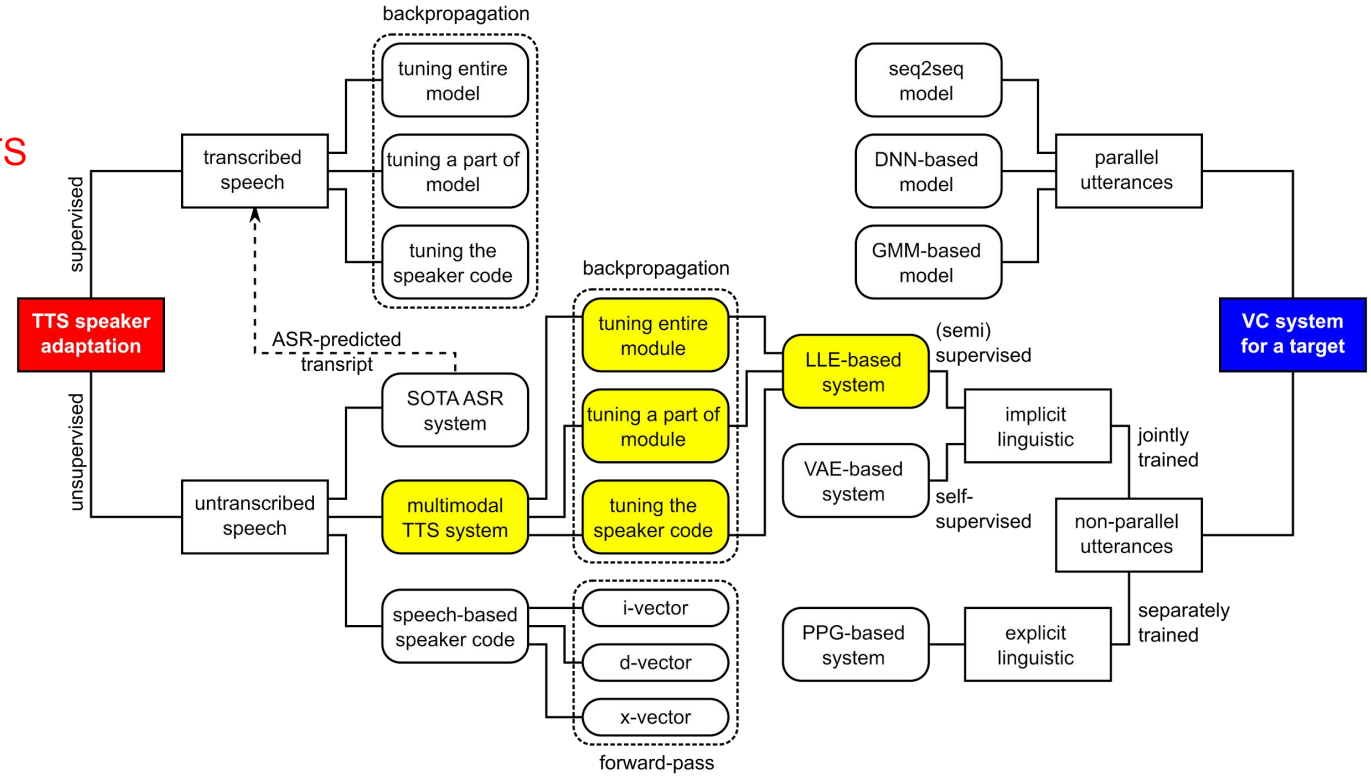
Part 3

A unified VC and TTS system

Chapter 8 of the thesis

TTS and VC voice cloning approaches

The same multimodal system has shown to be able to use for cloning voices with TTS and VC.



Next we deliver a SOTA unified TTS/VC system.

Components of modern TTS/VC system

Modern speech synthesis systems uses many different complementary techniques to generate high-quality synthetic speech.

Huang, Wen-Chin, et al. "Refined WaveNet vocoder for variational autoencoder based voice conversion." *Proc. EUSIPCO*, 2019.

Wang, Xin, Shinji Takaki, and Junichi Yamagishi. "An autoregressive recurrent mixture density network for parametric speech synthesis." *Proc. ICASSP*, 2017.

Watts, Oliver, et al. "Where do the improvements come from in sequence-to-sequence neural TTS?." *Proc. SSW*, 2019.

1. **Jointly tuned neural vocoder**

A neural vocoder that generates speech sample-by-sample (e.g. WaveNet) is important. It can be jointly tuned with the acoustic model using backpropagation as well.

2. **Autoregressive generation**

Autoregressive neural acoustic model, which generates current based on previous generated frames, is reported to improve naturalness of synthetic speech.

3. **Data-driven linguistic representation**

Instead of using hand-crafted linguistic features extracted using a language-dependent front-end, we let the neural network learn the relevant linguistic context on its own.

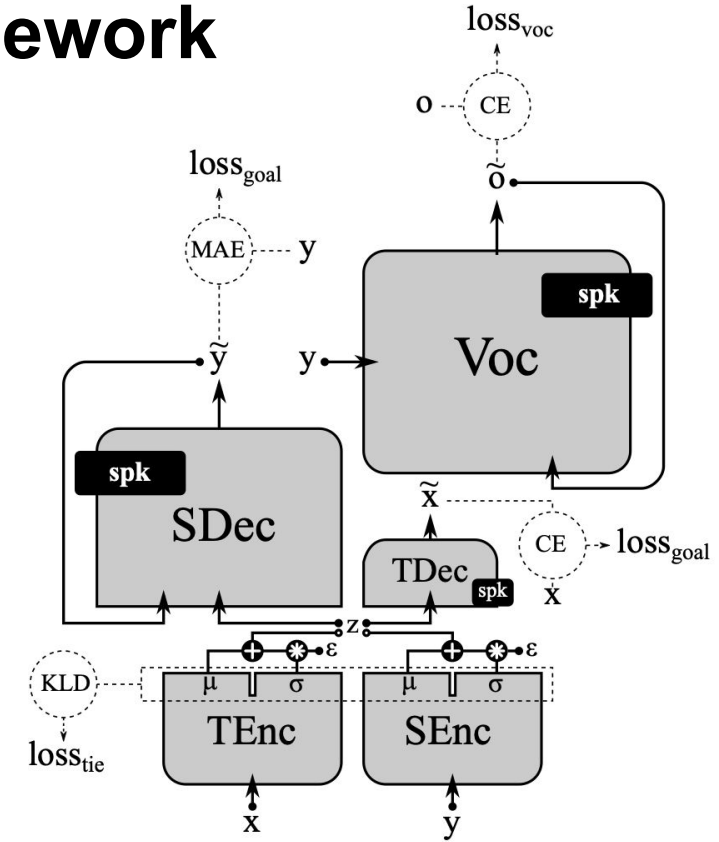
A versatile voice cloning framework

The **variational multimodal neural TTS + text decoder** used as auxiliary regularization. The model is trained with joint-goal and tied-layer:

$$\begin{aligned}\text{loss}_{train} &= \text{loss}_{goal} + \beta\ \text{loss}_{tie} \\ &= \text{loss}_{tts} + \alpha_{sts}\ \text{loss}_{sts} + \alpha_{stt}\ \text{loss}_{stt} \\ &\quad + \beta\ \text{loss}_{tie}\ ,\end{aligned}$$

using symmetric KL divergence as tied-layer loss:

$$\text{loss}_{tie} = \frac{1}{2}\ L_{KLD}(TEnc(\boldsymbol{x}), SEnc(\boldsymbol{y})) + \frac{1}{2}\ L_{KLD}(SEnc(\boldsymbol{y}), TEnc(\boldsymbol{x}))$$



Related work: semi-supervised jointly trained ASR and TTS

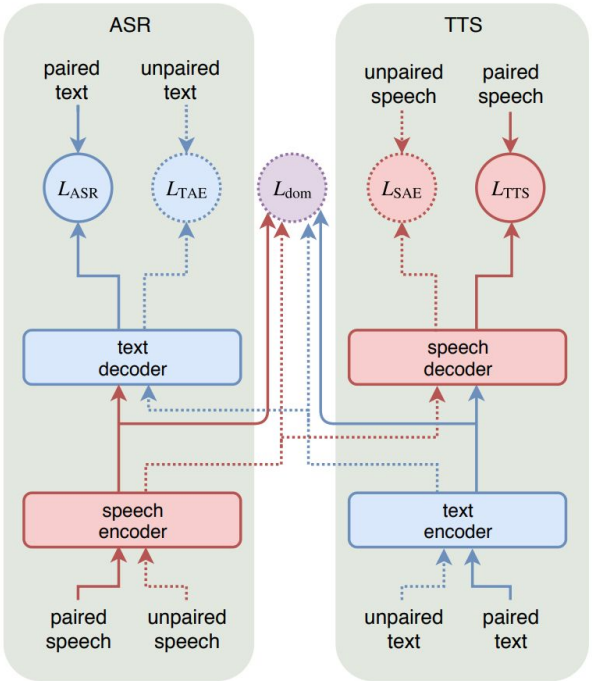
Exploit the ASR/TTS multimodal structure for semi-supervised training

Similarities:

- Using the text-speech multimodal neural structure.
- Using *joint-goal* and *tied-layer* to jointly train the model.

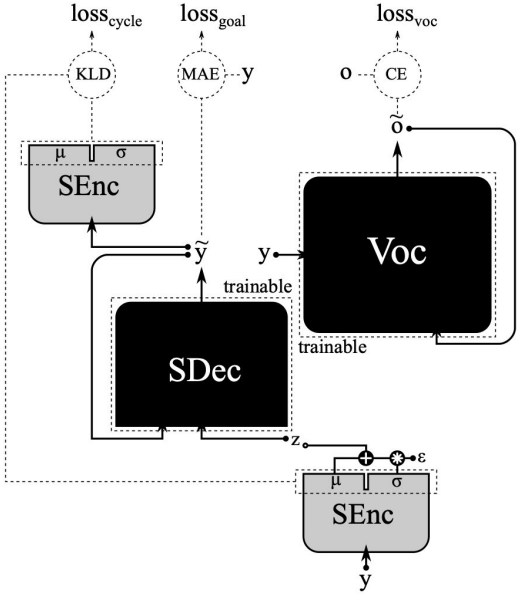
Differences:

- semi-supervised training (supervised training in our case).
- The text-speech multimodal is the end system (it is just the initial for adaptation in our case).
- The TTS-like system is only used as a leverage for ASR and not evaluated. (vice versa in our case)
- Technical details (VAE-like structure, speaker component,...)



Karita, Shigeki, et al. "Semi-supervised end-to-end speech recognition using text-to-speech and autoencoders." Proc. *ICASSP,* 2019.

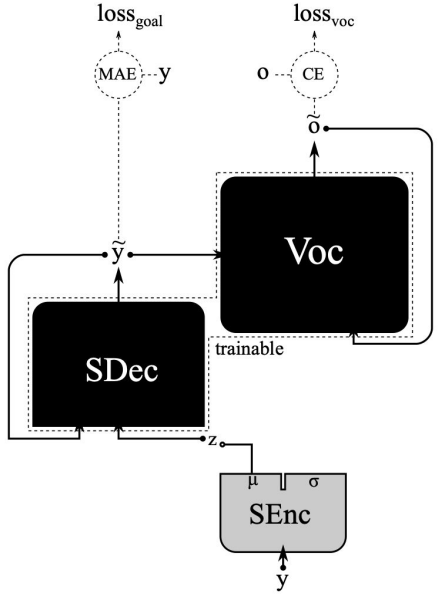
Cloning voices with untranscribed speech



Adaptation $\text{loss}_{adapt} = \text{loss}_{sts} + \beta\ \text{loss}_{cycle}$

STT loss and linguistic cycle consistent loss

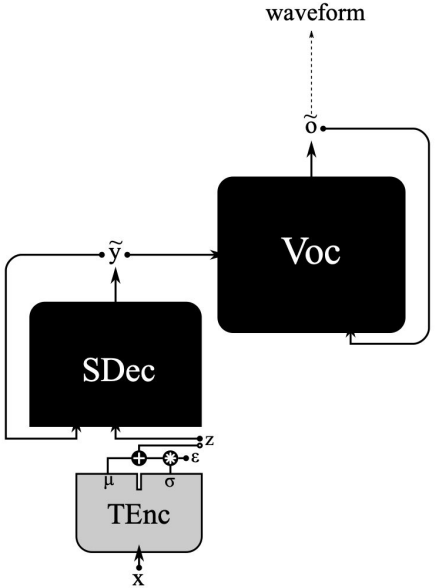
$$\text{loss}_{cycle} = \frac{1}{2} L_{KLD}(SEnc(\boldsymbol{y}), SEnc(\tilde{\boldsymbol{y}})) + \frac{1}{2} L_{KLD}(SEnc(\tilde{\boldsymbol{y}}), SEnc(\boldsymbol{y}))$$



Welding $\text{loss}_{weld} = \text{loss}_{sts} + \gamma\ \text{loss}_{voc}$

Jointly tuning speech decoder and neural vocoder to target speaker.

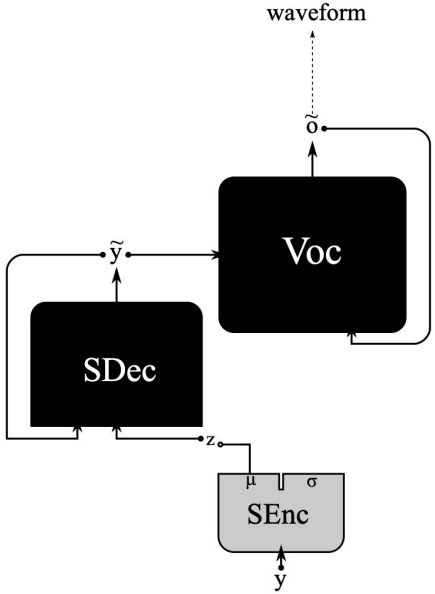
TTS/VC inferences



Inference/TTS

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

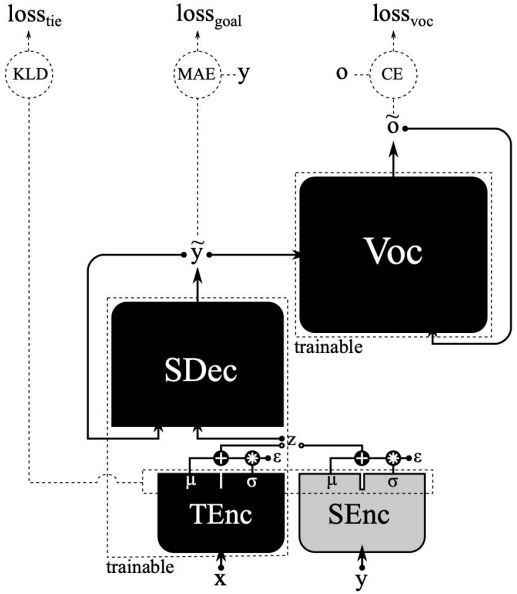
The same adapted **speech decoder** and **neural vocoder** are shared between two tasks.



Inference/VC

Converting speech of a source speaker to voice of a target speaker

Alternative strategy with transcribed speech



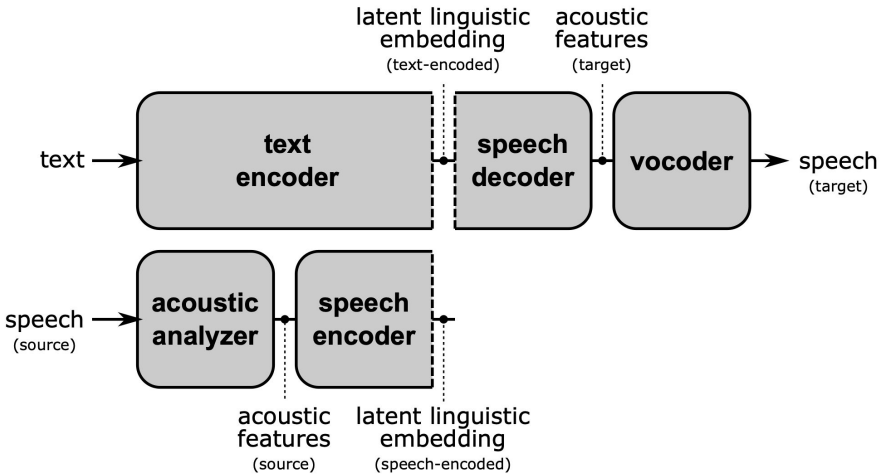
Adaptation
(supervised alternative)

Depending on data circumstance, we can adjust the cloning strategy to alter the behavior of TTS or VC system.

For example, using transcribed speech to tune both **text encoder** and **speech decoder** in adaptation stage:

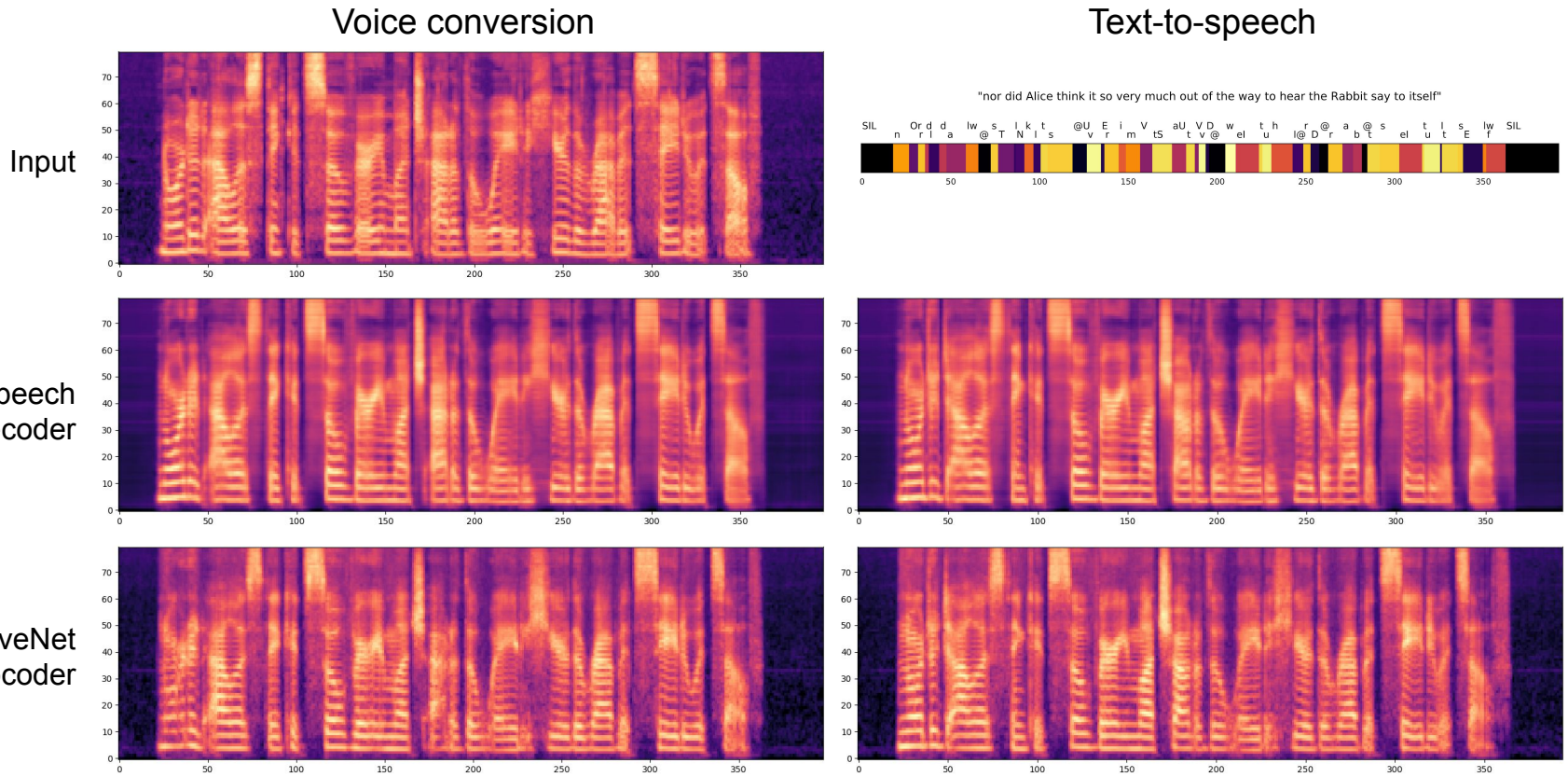
$$\text{loss}_{adapt} = \text{loss}_{tts} + \alpha\ \text{loss}_{sts} + \beta\ \text{loss}_{tie}$$

Pipeline of proposed unified TTS/VC system

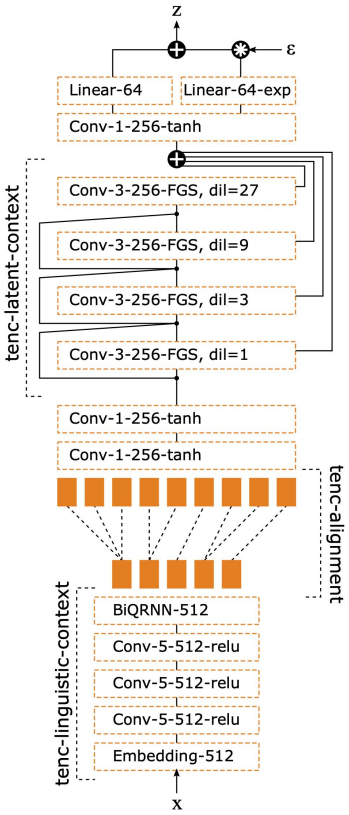


The performance consistency between TTS and VC of the proposed system is dependent on the consistency of the **text-encoded** and **speech-encoded** linguistic latent spaces.

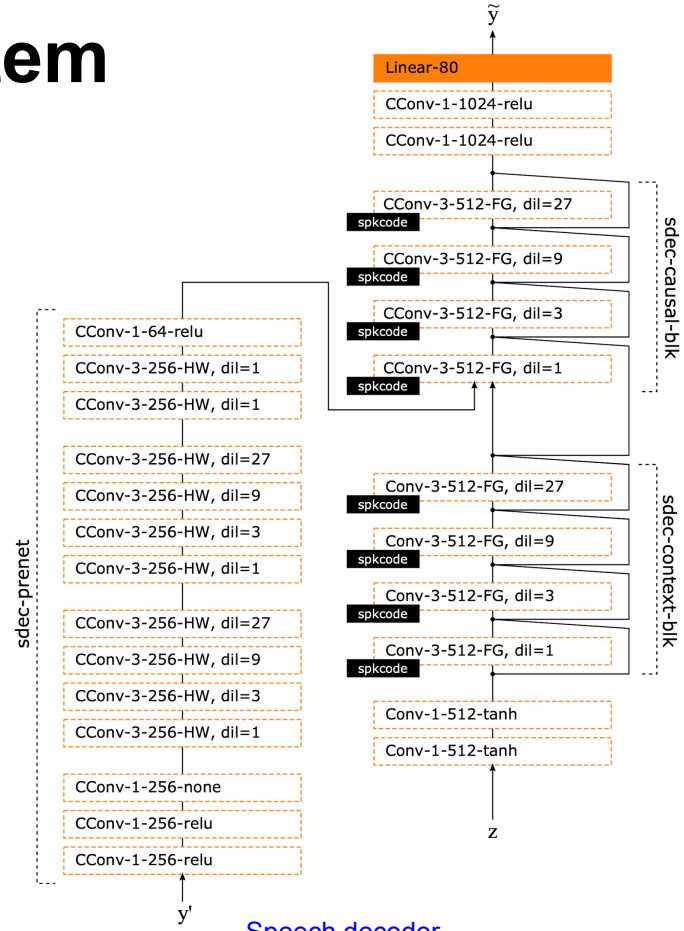
Pipeline demonstration



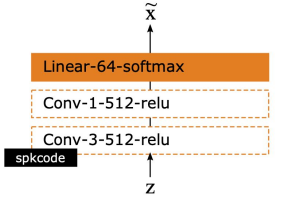
NAUTILUS system



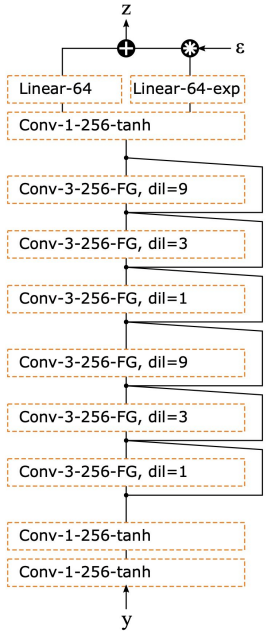
Text encoder



Speech decoder



Text decoder



Speech encoder

SCENARIO A - Description

Cloning voices using untranscribed speech

Evaluating the proposed system on the voice cloning task of both TTS and VC systems using a small amount (5 minutes) of untranscribed speech.

Same-gender (=) and cross-gender (x) voice conversion are treated as two separate entities for better understanding.

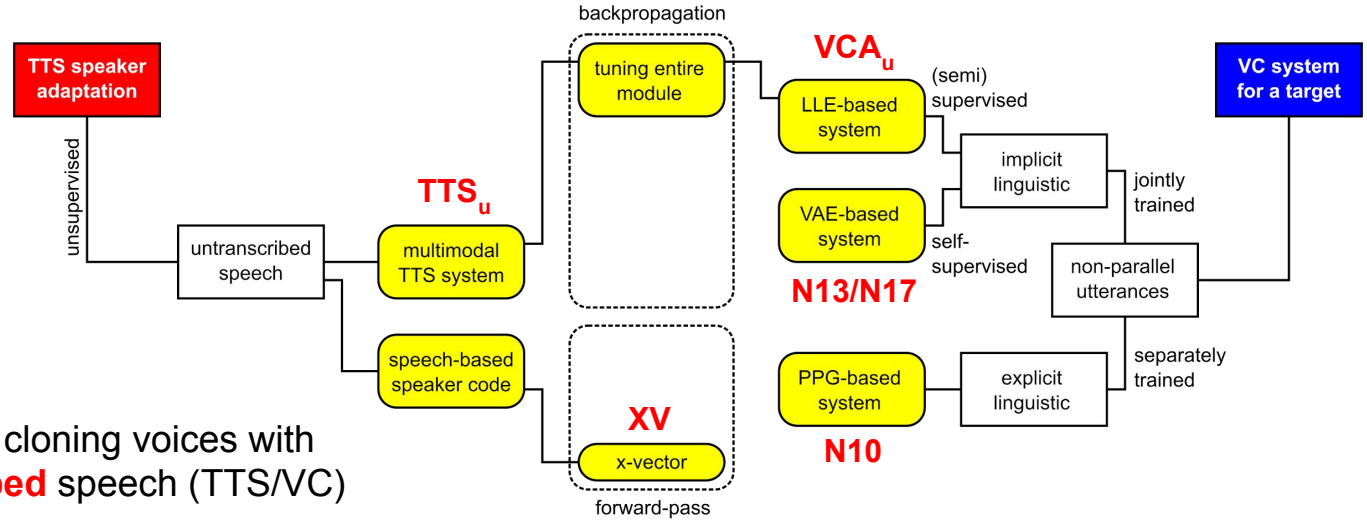
Table 8.1: Target speakers of scenario A.

Speaker	Database	Gender	Accent	Quantity	Duration
VCC2TF1	VCC2018	female	American	81 utt.	5.2 min
VCC2TF2	VCC2018	female	American	81 utt.	5.0 min
VCC2TM1	VCC2018	male	American	81 utt.	5.2 min
VCC2TM2	VCC2018	male	American	81 utt.	5.3 min

Reenact **VCC2018 SPOKE task** one more time but adding several extra TTS systems.

SCENARIO A - Systems

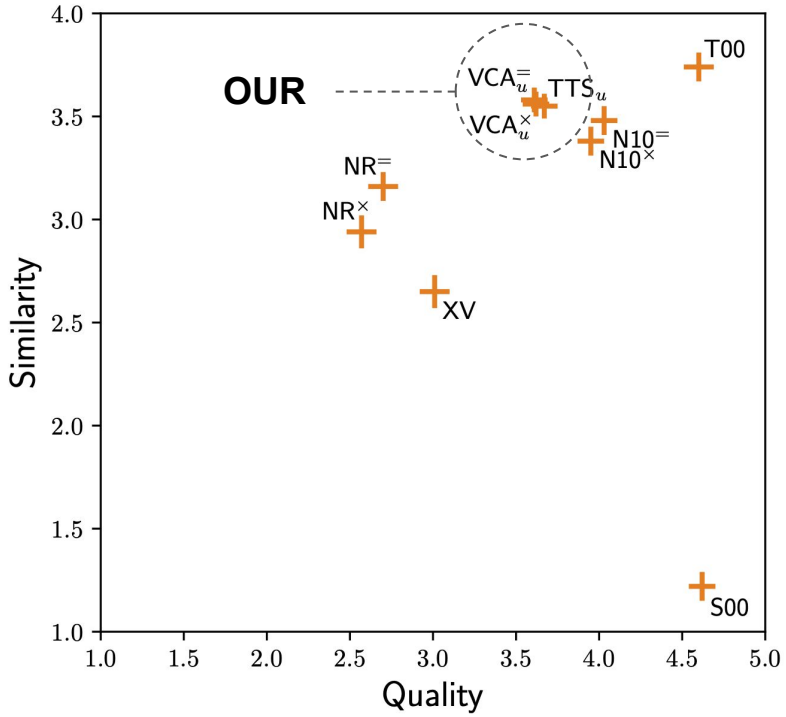
- **XV**: E2E TTS, unsupervised, x-vector
- **N10**: PPG-based VC, best of VCC2018
- **N13/N17 (NR)**: VAE/GAN-based, runner-up
- **VCA_u**: proposed VC, unsupervised
- **TTS_u**: proposed TTS, unsupervised



Focused on cloning voices with **untranscribed** speech (TTS/VC)

SCENARIO A - Evaluation

- Our system has slightly lower quality than **N10** but better similarity → SOTA performance consider all experiment condition mismatch (e.g., amount of training data)
- **VCA_u** and **TTS_u** have **consistent** quality and speaker similarity → useful for many application scenarios
- **XV** has decent quality and similarity → strong baseline.



Speech sample: https://nii-yamagishilab.github.io/sample-versatile-voice-cloning/

SCENARIO B - Description

Capture unique speaker characteristic

Can the voice cloning methods capture unique, subtle and local characteristic of a target speaker, whose voice is quite different from the training speakers.

This scenario focuses more on TTS and supervised/unsupervised speaker adaptation strategies.

Table 8.3: Target speakers of scenario B.

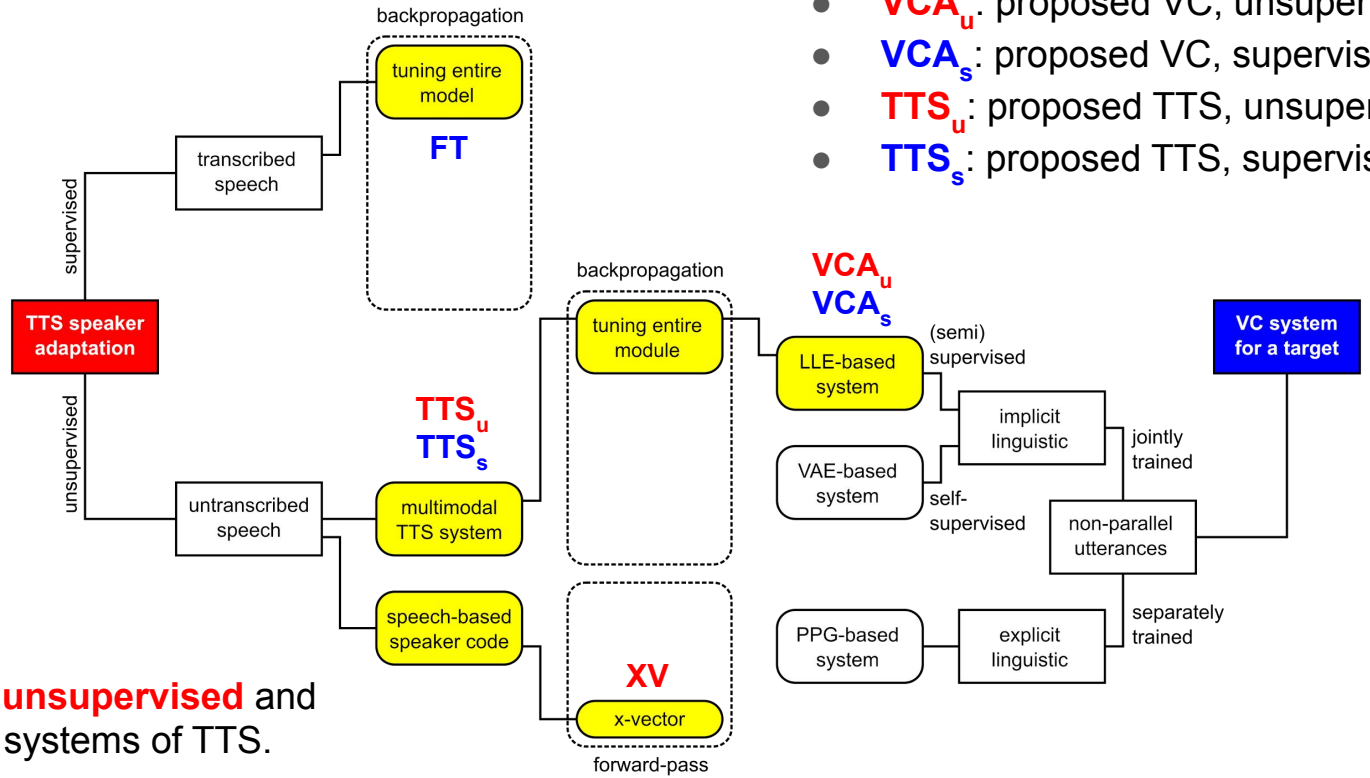
Speaker	Database	Gender	Accent/L1	Quantity	Duration
p294	VCTK	female	American	325 utt.	11.2 min
p345	VCTK	male	American	325 utt.	11.0 min
MF6	EMIME	female	Mandarin	145 utt.	10.2 min
MM6	EMIME	male	Mandarin	145 utt.	11.3 min

Two **American-accent** speakers are used as standard “easy” targets.
Two **Mandarin-L1** speakers are used as unique “difficult” targets.

The goal is cloning voices of target speakers and maintaining their unique accents.

SCENARIO B - Systems

- **XV**: E2E TTS, unsupervised, x-vector
- **FT**: E2E TTS, supervised, fine-tuning
- **VCA_u**: proposed VC, unsupervised
- **VCA_s**: proposed VC, supervised
- **TTS_u**: proposed TTS, unsupervised
- **TTS_s**: proposed TTS, supervised

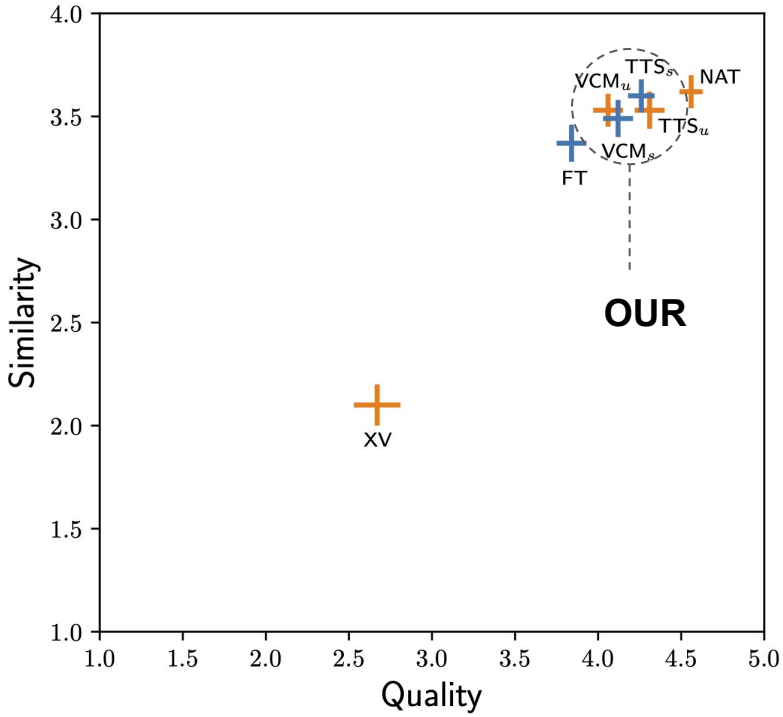


Focused on **unsupervised** and **supervised** systems of TTS.

SCENARIO B - Evaluation 1

Native (American) target speakers:

- Our TTS/VC systems and **FT** have high scores in quality and similarity. Both supervised and unsupervised strategies are better than supervised baseline, **FT**.
- No clear benefit of supervised adaptation strategy over unsupervised although **TTS_s** seems slightly better than **TTS_u** in term of speaker similarity
- The unsupervised baseline, **XV**, is not as good as the rest.



Most of the listener are English teachers lived in Japan.

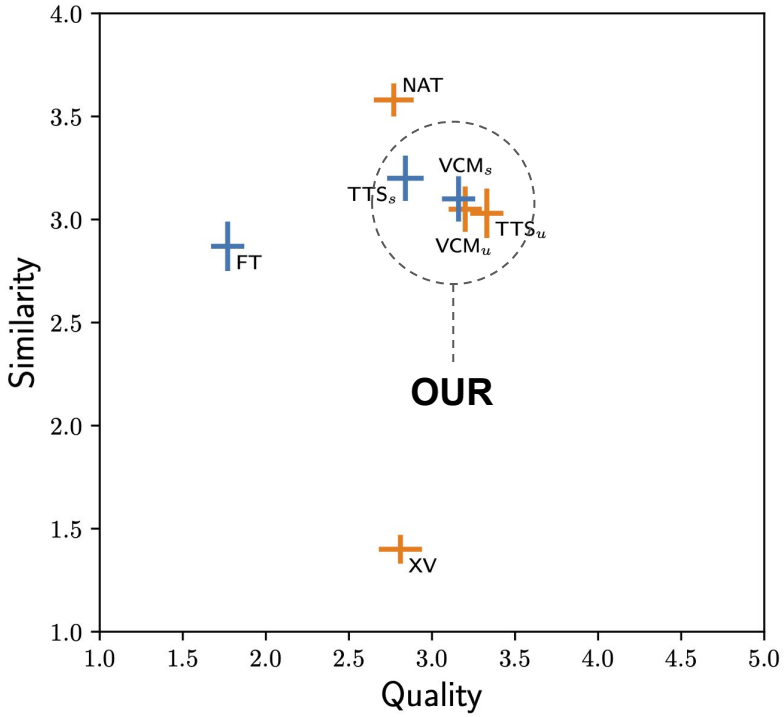
Speech sample: https://nii-yamagishilab.github.io/sample-versatile-voice-cloning/

SCENARIO B - Evaluation 2

Non-native (Mandarin-L1) target speakers:

- Natural speech **NAT** has low quality but maintains high speaker similarity.
- The supervised and unsupervised baseline (**FT** and **XV**) have opposite results in quality and similarity → negative correlation.
- **TTS_u** **VCA_u** **VCA_s** have better quality than natural speech → reduced accent?
- **TTS_s** has most similar evaluation to **NAT**.

XV, speaker codes based, is too conservative.
FT, fine-tuning based, is too liberal.
Our method is just right and offered two variations.



Most of the listener are English teachers lived in Japan.

Speech sample: https://nii-yamagishilab.github.io/sample-versatile-voice-cloning/

Conclusion

Contributions/Summary

FOR TTS RESEARCH

Proposed a novel multimodal neural TTS system which is capable of cloning voices with both transcribed and untranscribed speech with varying amount of adaptation data.

FOR VC RESEARCH

Proposed a novel framework for a robust non-parallel/cross-lingual VC bootstrapped from TTS which beats SOTA VC system on the speaker similarity measurement.

FOR THE NEW UNIFIED TTS/VC RESEARCH DIRECTION

Established the motivations and evaluations for a unified (or performance-consistent) TTS/VC system, and presented a strong system for such task.

Future works

CONTROLLABLE PARA-LINGUISTIC SPEECH GENERATION SYSTEM

Controlling linguistic content = speech synthesis, controlling speaker characteristic = voice cloning. Next we can enhance the speech generation system with ability to control other paralinguistic features (e.g., emotions) or increase the details of control over existing model (e.g., speaker accent)

MULTIMODAL SPEECH SYNTHESIS INTERFACES

TTS and VC can be seen as different input interfaces for generating speech. Given the multimodal structure of the proposed systems we can extend it to other less-common input interfaces (e.g., video-to-speech) and takes advantage of semi-supervised training.

Thank you for your attention

All the preprint papers and samples can be found at:
www.hieuthi.com