APSIPA ASC 2024@Macau Tutorial [T04] 15:30 ~ 17:30, Dec. 3, 2024.

Emerging Topics for Speech Synthesis: Versatility and Efficiency

Yuki Saito (University of Tokyo, Japan)
Shinnosuke Takamichi (Keio University, Japan)
Wataru Nakata (University of Tokyo, Japan)

Takamichi Laboratory
慶應義塾大学 高道研究室

MOONSHOT
RESEARCH & DEVELOPMENT PROGRAM

AVATAR SYMBIOTIC SOCIETY

Speaker information

- **Yuki Saito (齋藤 佑樹)**
  - Lecturer of University of Tokyo
  - Research interests: speech synthesis, voice conversion, etc.
  - Homepage: https://sython.org

- **Shinnosuke Takamichi (高道 慎之介)**
  - Associate Professor of Keio University
  - Research interests: speech processing, corpus, etc.
  - Homepage: https://takamichi-lab.github.io

- **Wataru Nakata (中田 亘)**
  - Master’s student of University of Tokyo
  - Research interests: speech language model, neural audio codec
  - Homepage: https://wataru-nakata.github.io/

Our work published at APSIPA ASC 2024

- Two oral presentations
  - **W. Nakata**, T. Saeki, **Y. Saito**, **S. Takamichi**, H. Saruwatari, "NecoBERT: Self-Supervised Learning Model Trained by Masked Language Modeling on Rich Acoustic Features Derived from Neural Audio Codec," Dec. 4, 14:40 ~ 15:00, Meeting Room 9 (**Best Student Paper Competition**)
  - Y. Ishikawa, O. Take, T. Nakamura, N. Takamune, **Y. Saito**, **S. Takamichi**, H. Saruwatari, "Real-Time Noise Estimation for Lombard-Effect Speech Synthesis in Human–Avatar Dialogue Systems," Dec. 6, 10:00 ~ 12:00, Meeting Room 4 (Special Session: Acoustic Scene Analysis and Signal Enhancement Based on Advanced Signal Processing and Machine Learning)

Table of contents

- Emerging Topics for Speech Synthesis: Versatility and Efficiency
  - Goal: Understanding emerging topics in the speech synthesis field
- Contents (2 hours)
  - Part 1: Introduction (10 min.)
  - Part 2: Data (30 min.)
  - Part 3: Learning (30 min.)
  - Short break (10 min.)
  - Part 4: Evaluation (30 min.)
  - Part 5: Summary & Outlook (10 min.)
- Acknowledgement
  - We thank Dr. Takaaki Saeki for reviewing our tutorial material.

Part 1/5: Introduction (10 min.)

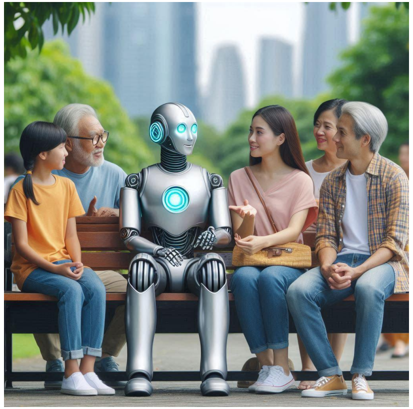
 Yuki Saito

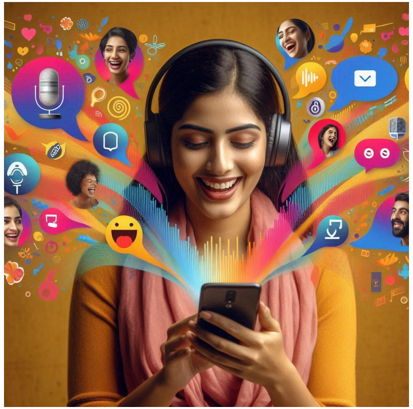
 Shinnosuke Takamichi

 Wataru Nakata

Background: **T**ext-**T**o-**S**peech

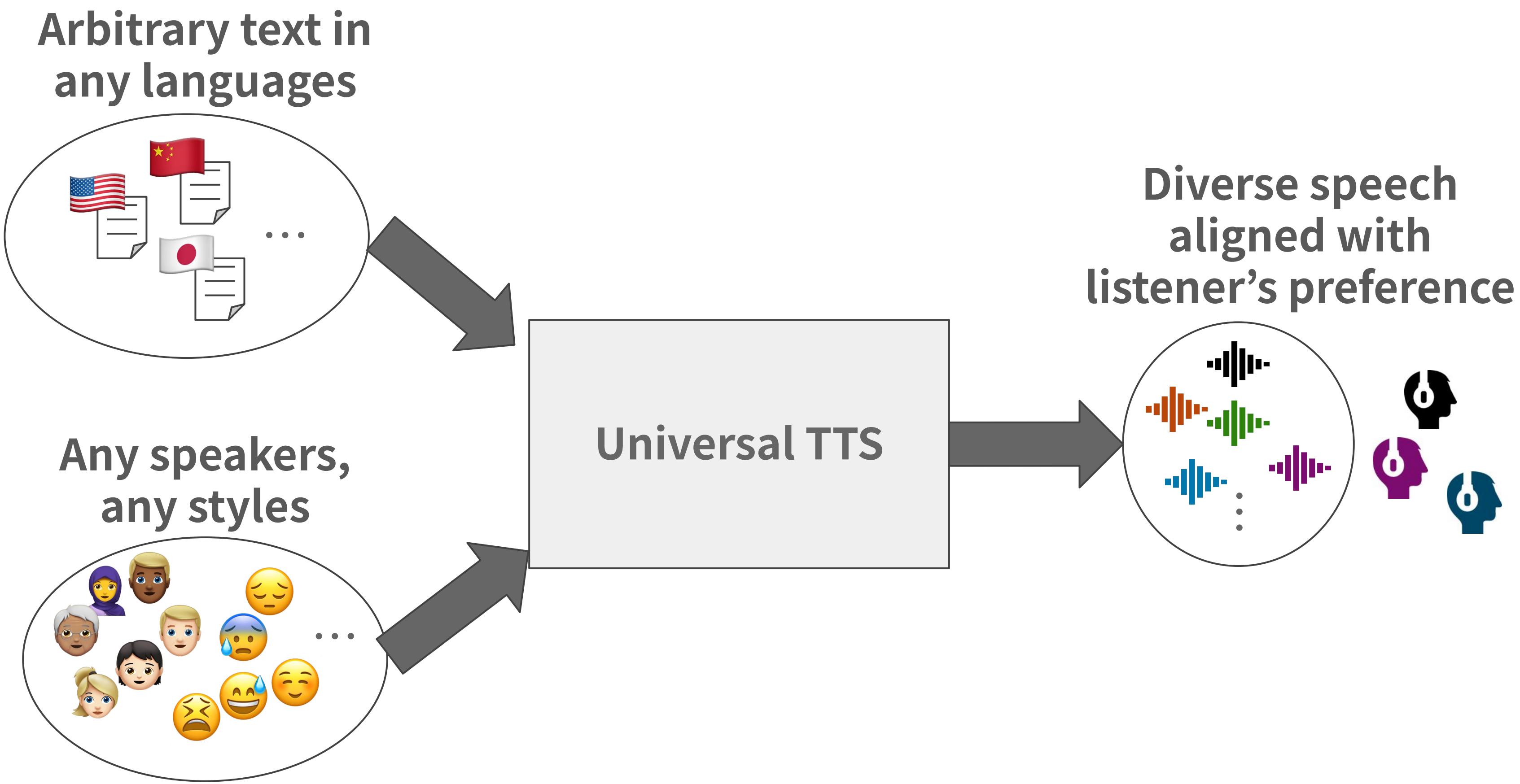
- Technology to artificially synthesize speech from given text
  - Enabling robots/computers to speak like humans
- Applications
  - Human machine interface
  - Speaking aids
  - Language education
  - Entertainments



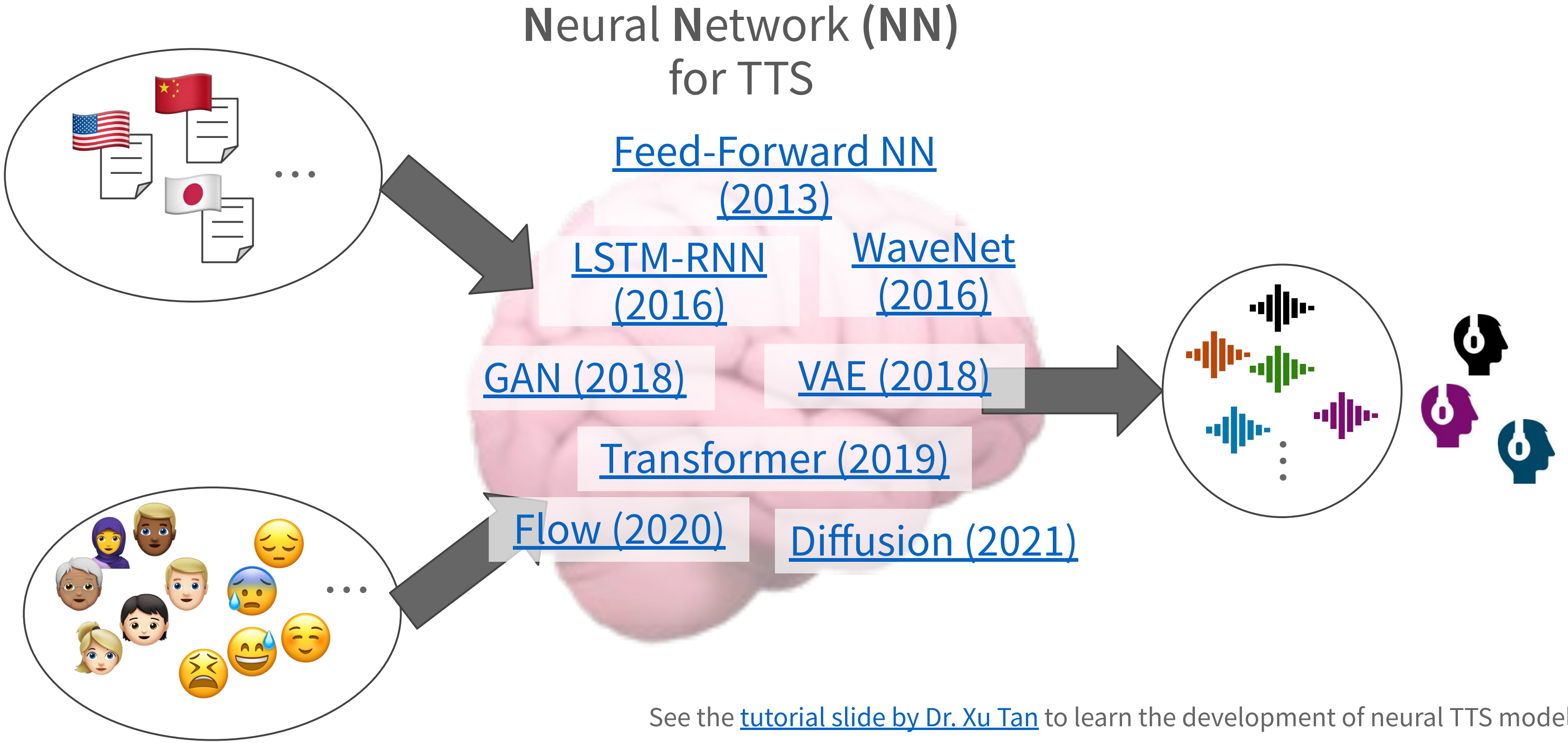


- c.f., **V**oice **C**onversion
  - Transforming characteristics of input speech into another ones
  - Both TTS & VC aims to naturally sounding speech
    → Knowledge & techniques can be shared within these technologies

Ultimate goal: building universal TTS model



Main topic of this tutorial: **neural TTS**



See the tutorial slide by Dr. Xu Tan to learn the development of neural TTS models.

Examples of state-of-the-art neural TTS technologies

Spoken dialogue system
[GPT-4o (2024)]

Zero-shot expressive TTS
[NaturalSpeech 3 (2024)]

Zero-shot multilingual TTS
[VALL-E X (2024)]

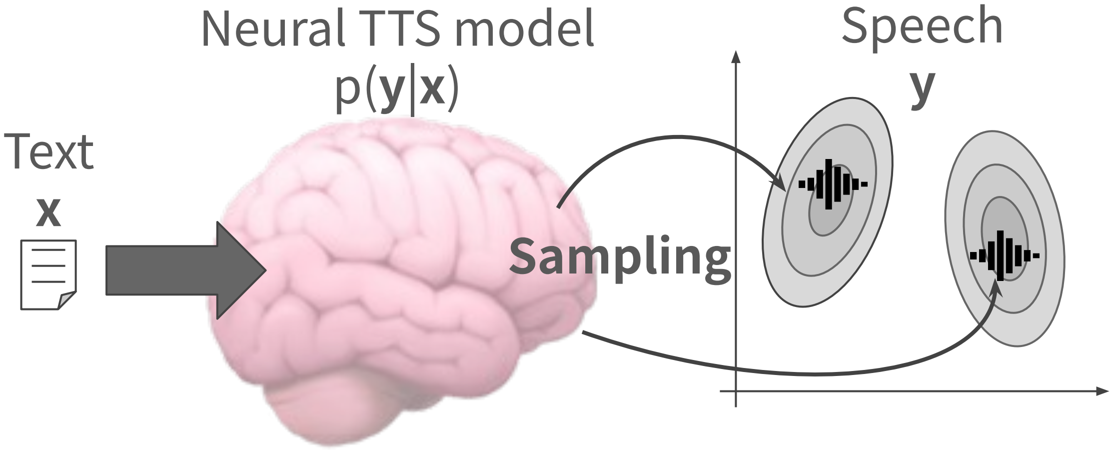
APSIPA is an emerging association to promote broad spectrum of research and education activities in signal and information processing.

APSIPA 是一个新兴协会，旨在促进信号和信息处理领域的广泛研究和教育活动。

The VALL-E X demo samples were generated by this HuggingFace page.

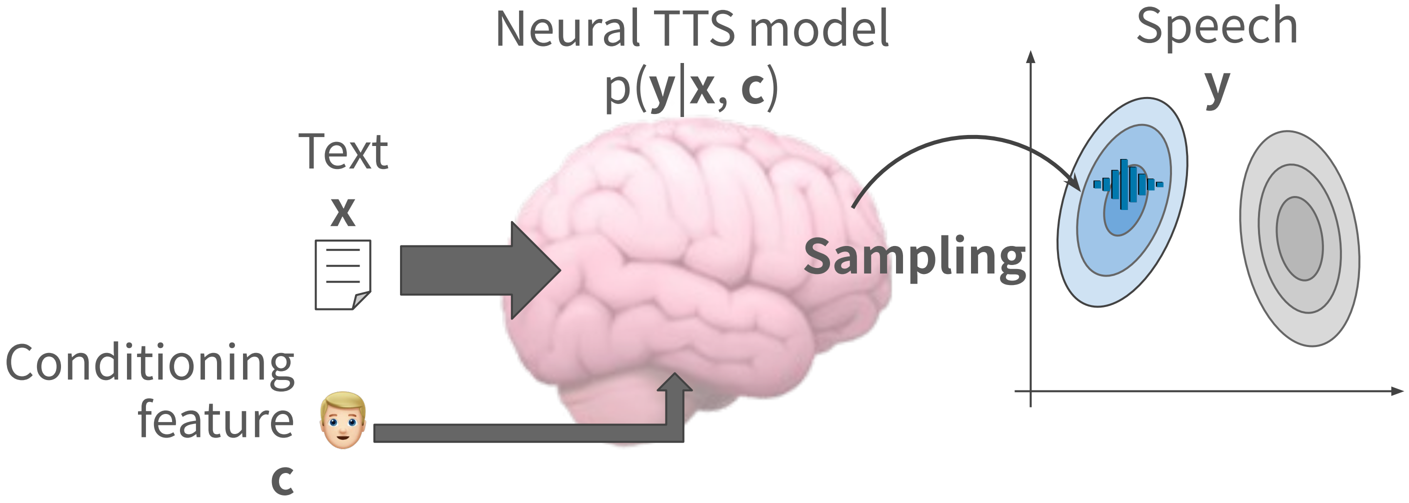
Probabilistic formulation of neural TTS

- Learning the probability distribution of speech conditioned on text
  - The distribution is **multimodal**, because one text can be spoken from any speakers with any speaking styles → **one-to-many mapping**!



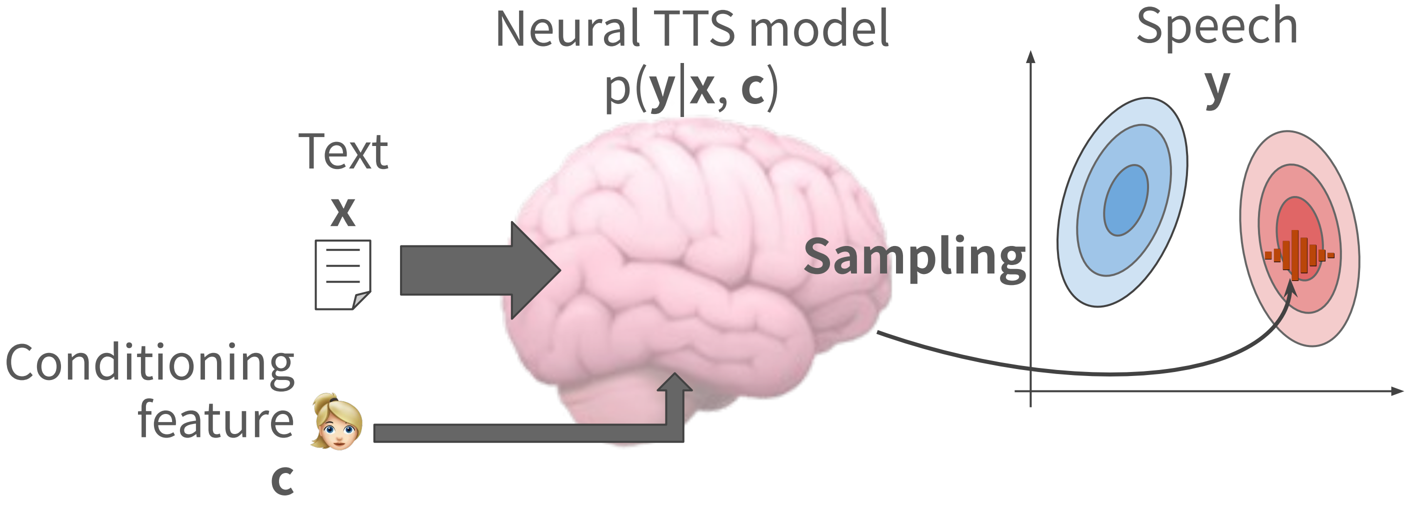
Probabilistic formulation of neural TTS

- Learning the probability distribution of speech conditioned on text
  - The distribution is **multimodal**, because one text can be spoken from any speakers with any speaking styles → **one-to-many mapping**!
  - To control the sampling process, we can also consider **conditioning feature** as the input to the TTS model.



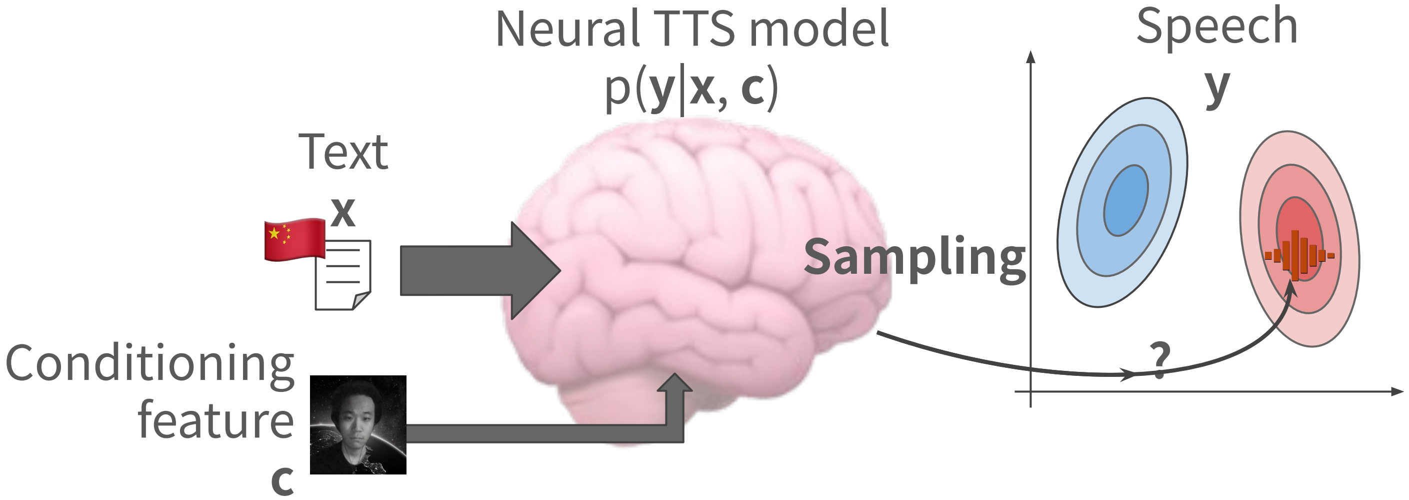
Probabilistic formulation of neural TTS

- Learning the probability distribution of speech conditioned on text
  - The distribution is **multimodal**, because one text can be spoken from any speakers with any speaking styles → **one-to-many mapping**!
  - To control the sampling process, we can also consider **conditioning feature** as the input to the TTS model.



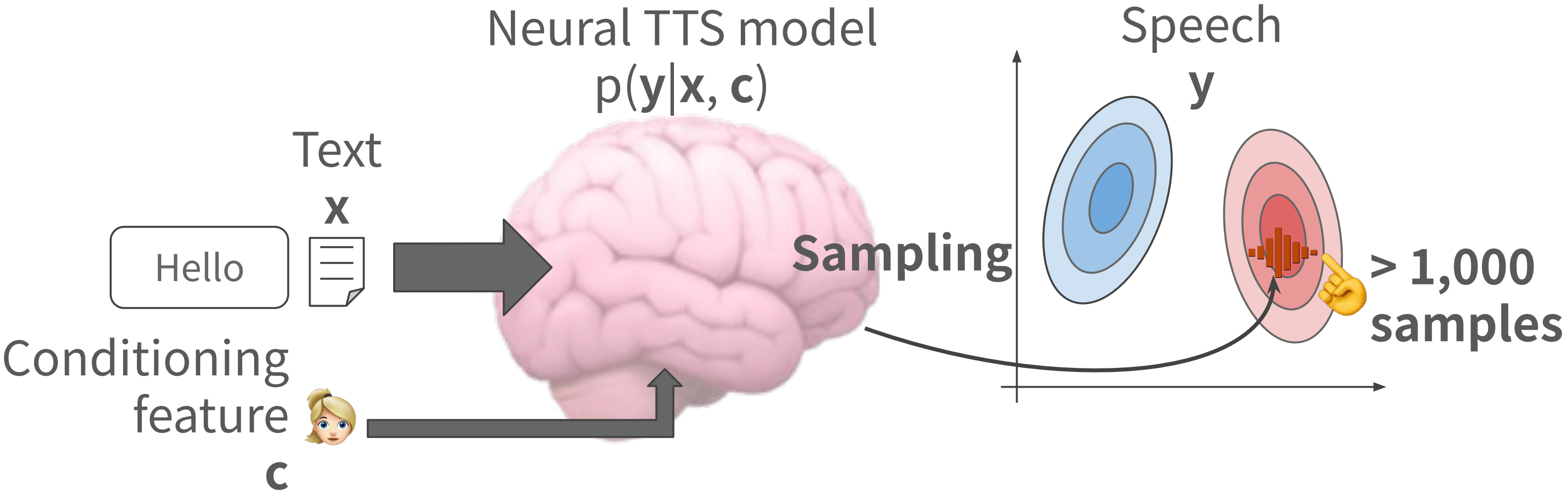
Essential factors for neural TTS (1/3): data

- Many **pairs** of (**x**, **y**, **c**) are required to learn p(**y**|**x**, **c**).
  - Collecting **x** or **y** is relatively easy, but (**x**, **y**, **c**) for all pairs is impossible.
  - Obtaining (**x**, **c**) from **y** requires laborious **annotations**.



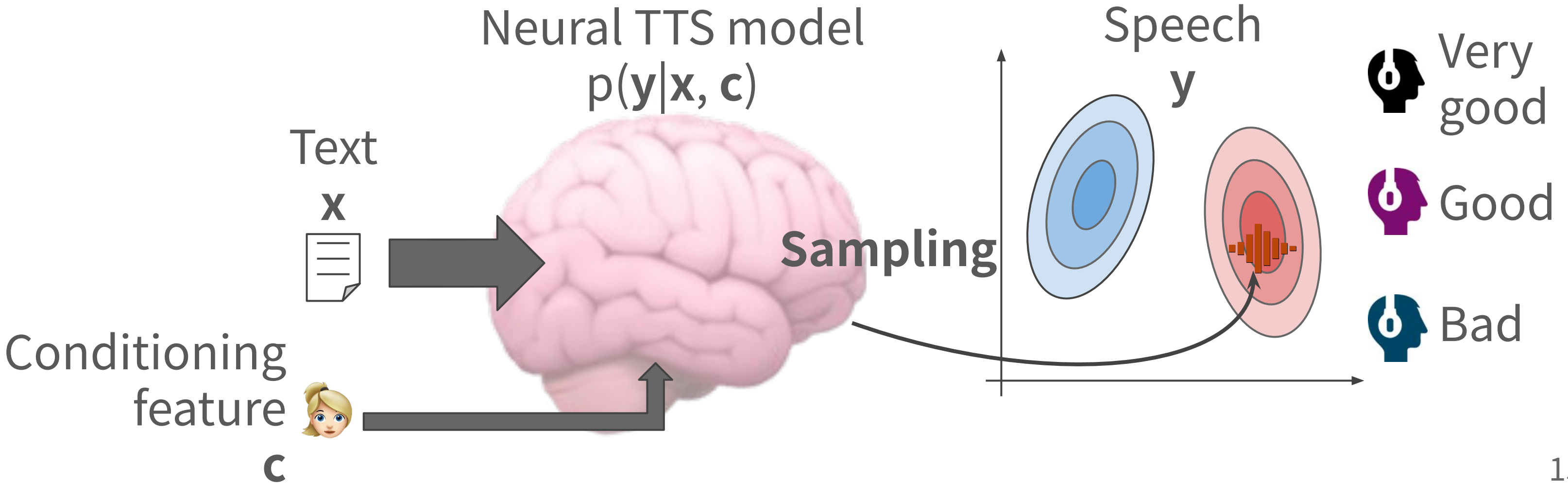
Essential factors for neural TTS (2/3): learning

- TTS can be regarded as sequence-to-sequence mapping problem.
  - **x**: character sequence, **y**: speech sample sequence (high density)
    - **c** can be also sequential data, depending on application.



Essential factors for neural TTS (3/3): evaluation

- Generated speech samples should be **accepted** by human listeners.
  - Human evaluation is intrinsically subjective and difficult to reproduce.
  - Objective evaluation is easy, but not correlated with human evaluation.



Summary of Part 1: introduction

- TTS: Technology to artificially synthesize speech from given text
  - Enabling robots/computers to speak like humans
- Neural TTS: using (**D**eep) NNs for learning the TTS mapping
  - Essential factors: data, learning, and evaluation
- Core research questions
  - **How can we effectively collect data for the universal TTS model?**
  - **How can we effectively learn/control TTS model?**
  - **How can we universally/fairly evaluate multiple TTS models?**

Through this tutorial, you will be able to get closer to the answers for these questions! 😄

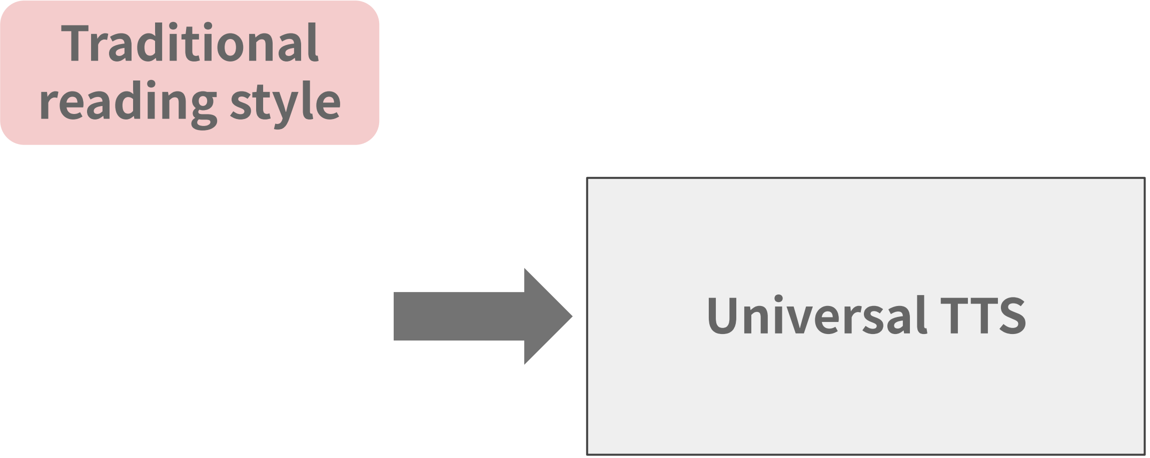
Part 2/5: Data (30 min.)

 Yuki Saito

 Shinnosuke Takamichi

 Wataru Nakata

Corpora for speech synthesis



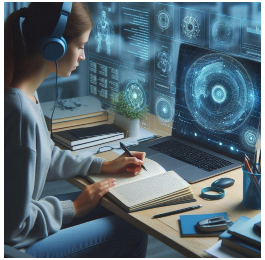
Dataset for traditional text-to-speech, i.e., reading-style TTS

Requirements for traditional TTS corpora

“Traditional” does not mean
“unnecessary in today.”

Controlled recording

- Speaker: professional, native, intelligible, etc.
- Environment: well-equipped, less noise, less reverberation



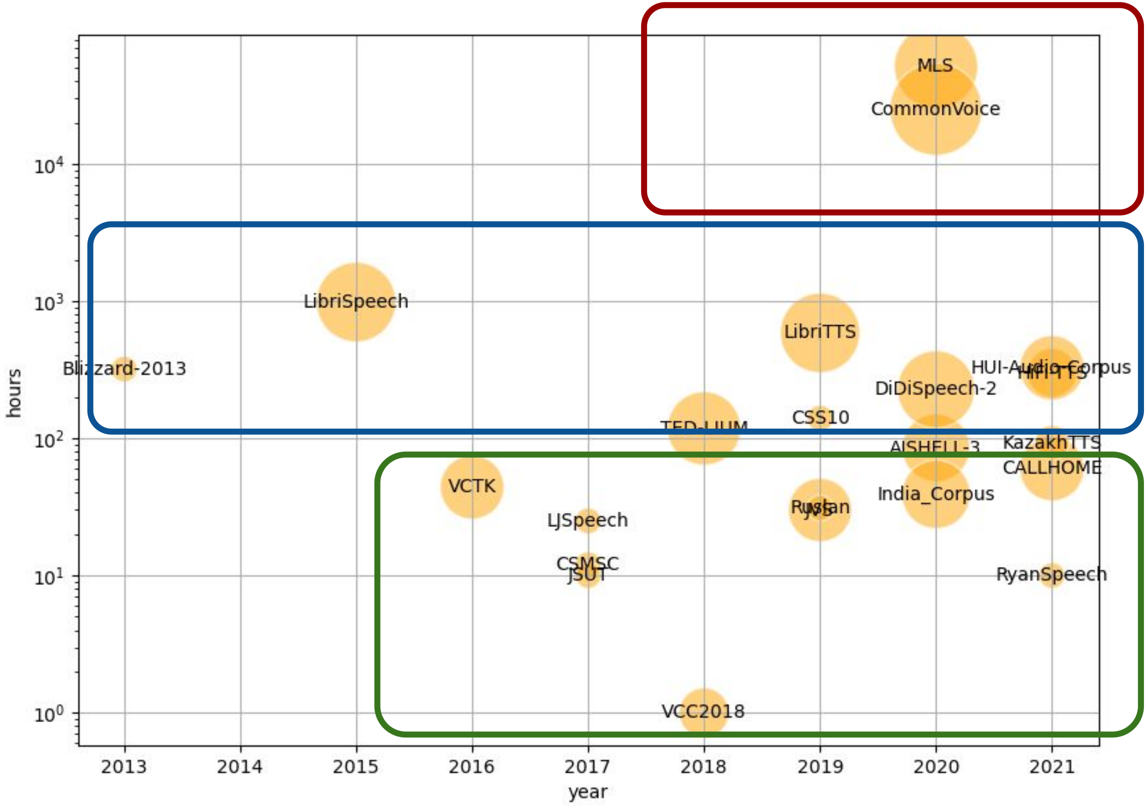
Rich annotation

- Speech: transcription, emotion, disfluency, alignment
- Speaker & environment: gender, skill, studio

Mid-large size

- 10+ hours / speaker
- multiple speakers

Growth of traditional TTS corpora (-2021)



Large (1000+ hours):

- multiple language / corpus
- relatively low-quality

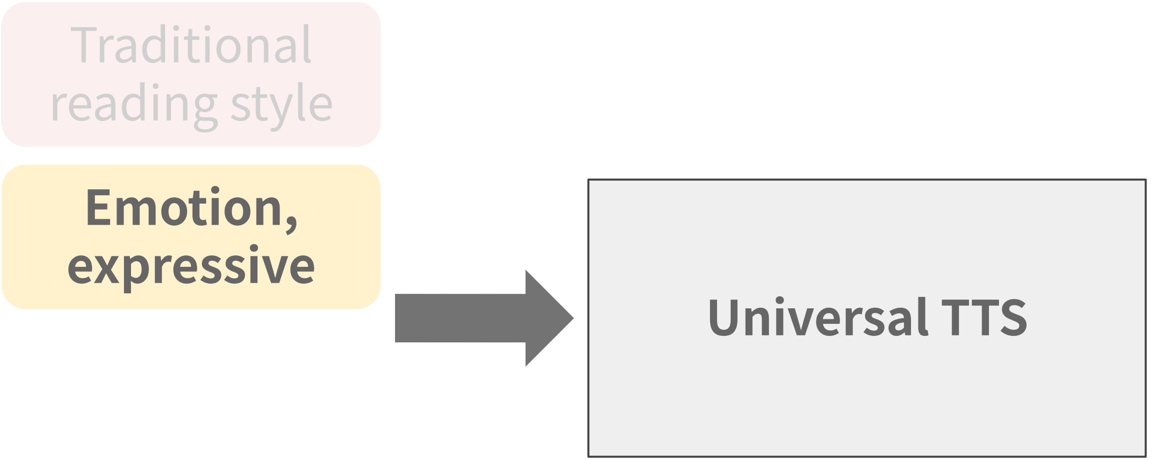
medium large (< 1000 hours):

- single language / corpus
- limited number of languages (en, zh, de, etc.)

medium (< 100 hours):

- single language / corpus
- relatively high-quality
- Many languages

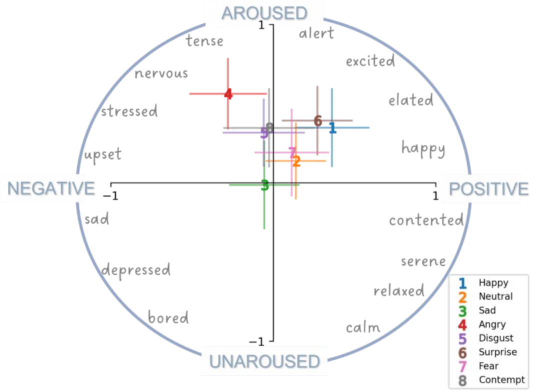
Corpora for speech synthesis



Let’s trancend traditional reading style!

Emotional data, expressive data

- Description of emotions
  - Categorical: happy, sad, etc.
    - + (discrete/continuous) intensity
  - Continuous: e.g., valence-arousal-dominance



https://www.microsoft.com/en-us/research/project/emoctrl-tts/

- Content type
  - Acted: the speaker creates and acts stereotypical emotion
  - Evoked: the speaker creates natural emotion under controlled condition.

- Environment
  - Studio (anechoic room) is preferable for acted emotion
  - Dailylife room is preferable for natural emotion

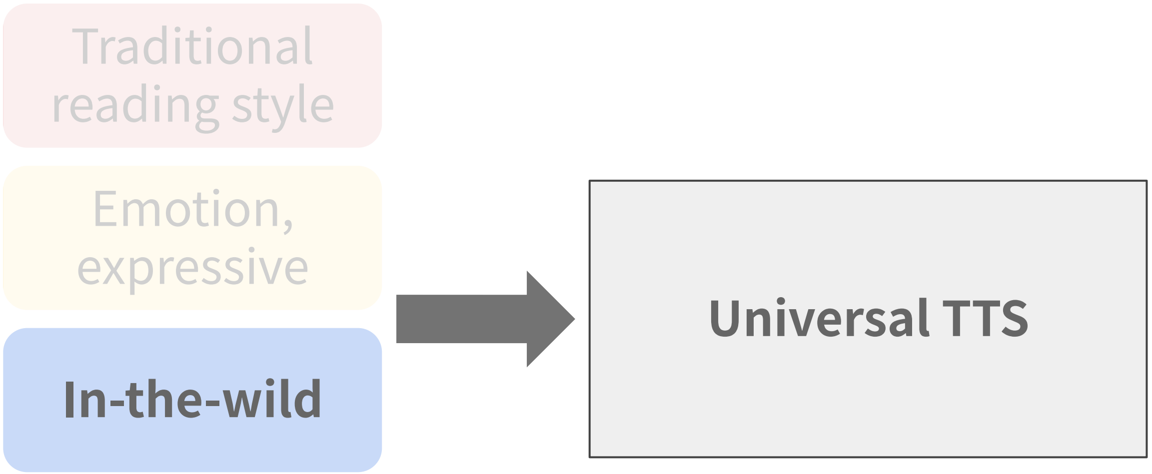
Recent direction of emotion corpus

- **Textual representation of emotion**
  - Automated caption:
    - “it sounds like this person is angry” [Liang24]
    - “high tone,” “low pitch”
  - Human-annotated caption:
    - “tone reveals inner dissatisfaction” [Xu24]

- **Non-verbal vocalization (NVV)**
  - Laughter: laughterscape [Xin23], AMI [Carletta07], DiariST-AliMeeting [Yang23]
  - Scream: Human Screaming Detection Dataset [whats2000_23]
  - Crying, sobs, etc.

[Xin23] https://arxiv.org/abs/2305.12442, [carletta07] https://www.openslr.org/16/, [Yang23] https://arxiv.org/abs/2309.08007, [whats2000_23] https://www.kaggle.com/datasets/whats2000/human-screaming-detection-dataset, [Liang24] https://arxiv.org/pdf/2410.19134, [Xu24] https://arxiv.org/pdf/2312.10381,

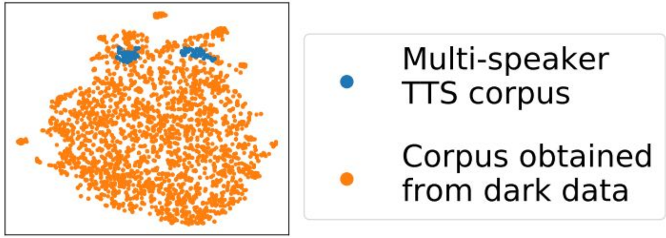
Corpora for speech synthesis



Is controlled recording enough for TTS?
-> No! Let's use in-the-wild data!

In-the-wild data;
is controlled recording NOT enough for TTS?

- Small in training data size:
  - Controlled recording is time-consuming work.
  - Larger data (+ larger model) cause emergent abilities in TTS [Łajszczak24].



https://arxiv.org/pdf/2210.14850

- Less diverse:
  - In speaker space: see right figure
  - In style and emotion spaces:
    - Difficult to record spontaneous style/emotion in controlled recording

- However…
  - In-the-wild data (e.g., web data) is not clean in various perspectives.
  - -> Data cleansing and data selection

[Łajszczak24] https://arxiv.org/abs/2402.08093,

Data cleasing

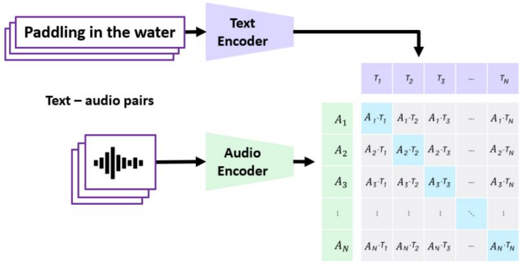
- Cascading data cleansing and model training
  - Past (before E2E era): errors in cleansing propagates to model training.
  - Nowadays: the cleansing quality got enough!
- Acoustic noise/distortion
  - AudioSR: super-resolution (e.g., 4kHz -> 48kHz)[Liu24]
  - Demucs v4: vocal-instrument separation [Defossez21]
  - Open-Miipher: speech restoration [Nakata24]
  - URGENT challenge: universal speech enhancement [Urgent24]
- Text noise
  - CTC: Re-segmentation of long audio [Kürzinger20]
  - Whisper v3: pre-trained ASR [Radford22]

[Liu24] https://arxiv.org/abs/2309.07314, [defossez21] https://github.com/facebookresearch/demucs, [Nakata24] https://github.com/Wataru-Nakata/miipher, [Urgent24] https://urgent-challenge.github.io/urgent2024/, [Kürzinger20] https://arxiv.org/abs/2007.09127, [Radford22] https://arxiv.org/abs/2212.04356,

Data selection by quantifying quality

- Audio quality (no need of reference audio)
  - DNSMOS: Automatic MOS prediction for noise suppression [Reddy20]
  - NISQA: Automatic MOS prediction for recording quality [Mittag21]
  - Torchaudio-squim: ML-based non-instructive STOI and PESQ [Anurag23]

- Text quality (no need of reference audio)
  - MLMscore: LLM likelihood [Salazar19]
  - GPTscore: LLM likelihood w/ prompts [Fu23]

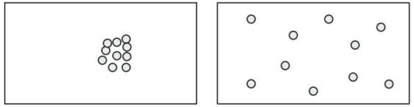


- Text-audio alignment quality
  - ASR score (CTC score [Kurzinger20], Whisper likelihood [Radford22])
  - Cosine dist of contrastive learning (e.g., CLAP) [Elizalde22]

[watanabe24] https://arxiv.org/abs/2403.13353, [Elizalde22] https://arxiv.org/abs/2206.04769, [Kürzinger20] https://arxiv.org/abs/2007.09127, [Radford22] https://arxiv.org/abs/2212.04356, [Sinha21] https://arxiv.org/abs/2104.06644, [Salazar19] https://arxiv.org/abs/1910.14659, [Fu23] https://arxiv.org/abs/2302.04166, [Reddy20] https://arxiv.org/abs/2010.15258, [Mittag21] https://arxiv.org/abs/2104.09494, [Anurag23] https://arxiv.org/abs/2304.01448,

Data selection (advanced)

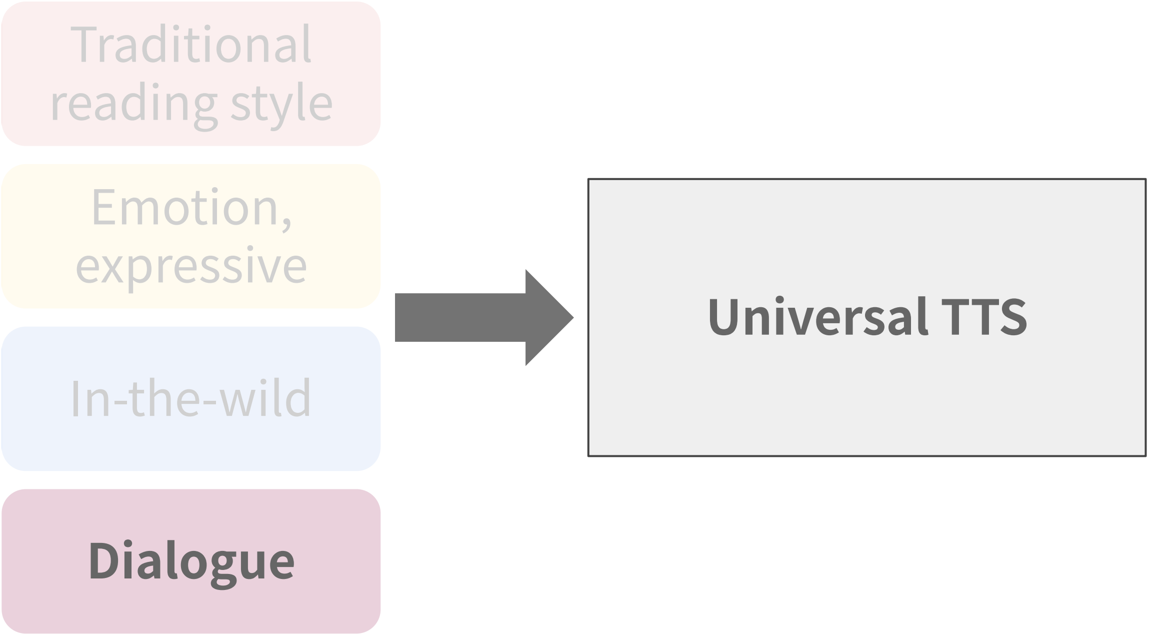
- Data duplication detection
  - Copy detection by SSL: avoiding unexpected data bias/leak [Pham23]
- Audio deepfake detection
  - ASVspoof: ASV and spoofing countermeasures [asvspoof]
  - ADD: audio deeepfake detection [ADD]
- Data uniformity and diversity
  - maximize uniformity: environment vector [Gallegos20], speaker vector [Takamichi21]
  - maximize diversity: BERT feature, SSL feature [Seki24]



(b) The core-set aims to cover an entire range, not a specific region.

[Pham23] https://www.isca-archive.org/interspeech_2023/pham23b_interspeech.html, [ASVspoof] https://www.asvspoof.org/, [ADD] http://addchallenge.cn/, [Gallegos20] https://www.isca-archive.org/interspeech_2020/gallegos20_interspeech.html, [Seki24] https://arxiv.org/abs/2309.08127,

Corpora for speech synthesis



Recent hot topic (chatGPT voice mode, Kyutai Moshi, etc.)!

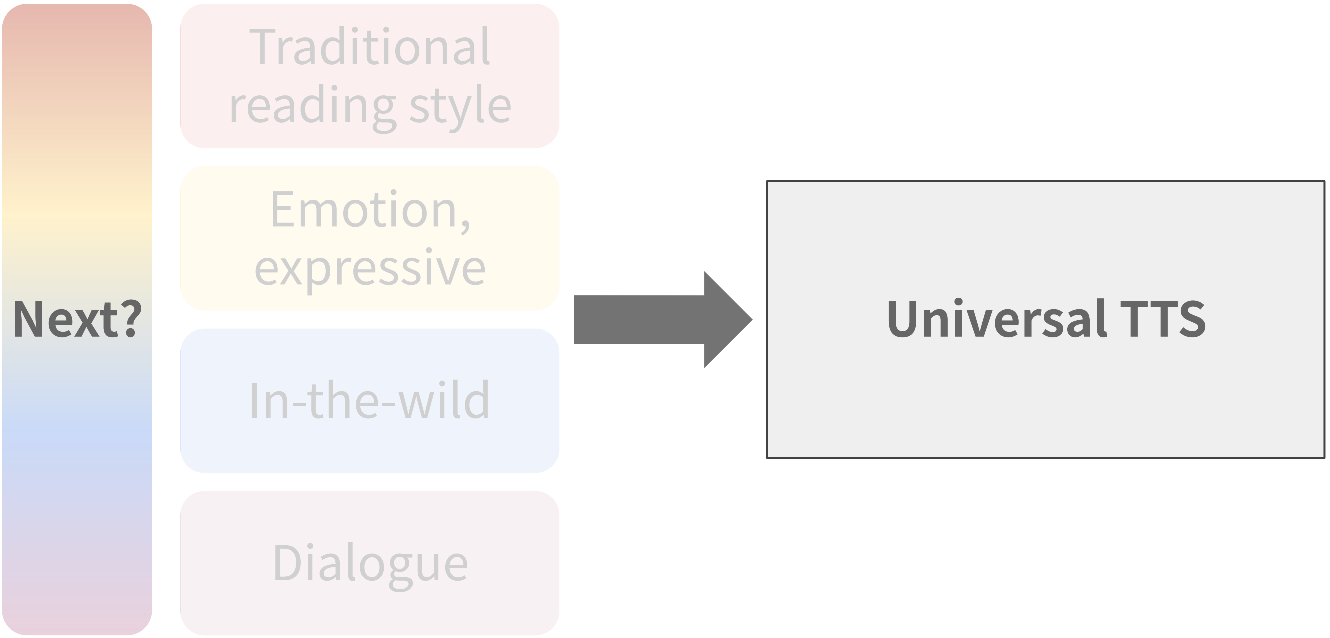
Dialogue

- Requirements & preferred
  - Multi-turn conversation
  - Speaker-separate audio channels
  - Backchannel (lexical and non-lexical)
  - Annotation of interpausal unit, speaker, etc.

- Very limited number of open-source corpora…
  - Ami [Carletta07] (0.1k hours, En)
  - ALLIES [Tahon24] (0.5k hours, Fr),
  - Fisher [Cieri04] (2k hours, En)
  - J-CHAT [Wataru24] (70k hours, Ja)
  - Finding in-the-wild dialogue data using dialogue consistency [Cekic22]

[Tahon24] https://aclanthology.org/2024.lrec-main.67/, [carletta07] https://www.openslr.org/16/, [Cieri04] https://catalog.ldc.upenn.edu/LDC2004T19, [Wataru24] https://arxiv.org/abs/2407.15828, [Cekic22] https://arxiv.org/abs/2202.03484,

Corpora for speech synthesis



What are waiting for us in the future?

What’s next?

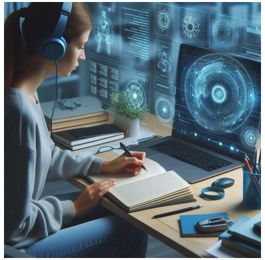
- More-complicated
  - Non-{visual, textual} prompts that describe speech (e.g., unconscious).
- Multi-modal
  - Non-speech audio, animal speech, vision, haptics, muscle, smell, etc.
- Artificial data
  - Real-person data has ethical issue.
- Data for unlearning [Zhang24]
  - Data to let a model not reproduce, e.g., important person.

[Zhang24] https://arxiv.org/abs/2411.11315,

What's next?: (currently) no speech data

Historical (e.g., earlier than 1900s)

- Deceased person
- Extinct languages



Regional & Social

- Language variants (e.g., dialects)
- Social & individual variants (e.g., diglossia, idiolect)

Fictional & artificial

- Fictional language in games
- Artificial language (one trend in NLP [Artetxe20])

[Artetxe20] https://aclanthology.org/2020.acl-main.421.pdf

Summary of part 2: data

- Controlled recording is traditional but still important.
- Description way is shifting from labels to free-form description.
- In-the-wild data is promising for TTS. (ethical issues remain.)
- We need to try more complicated tasks, e.g., unconscious.
- Many domain of data (or no data) are waiting for us.

Part 3/5: Learning (30 min.)

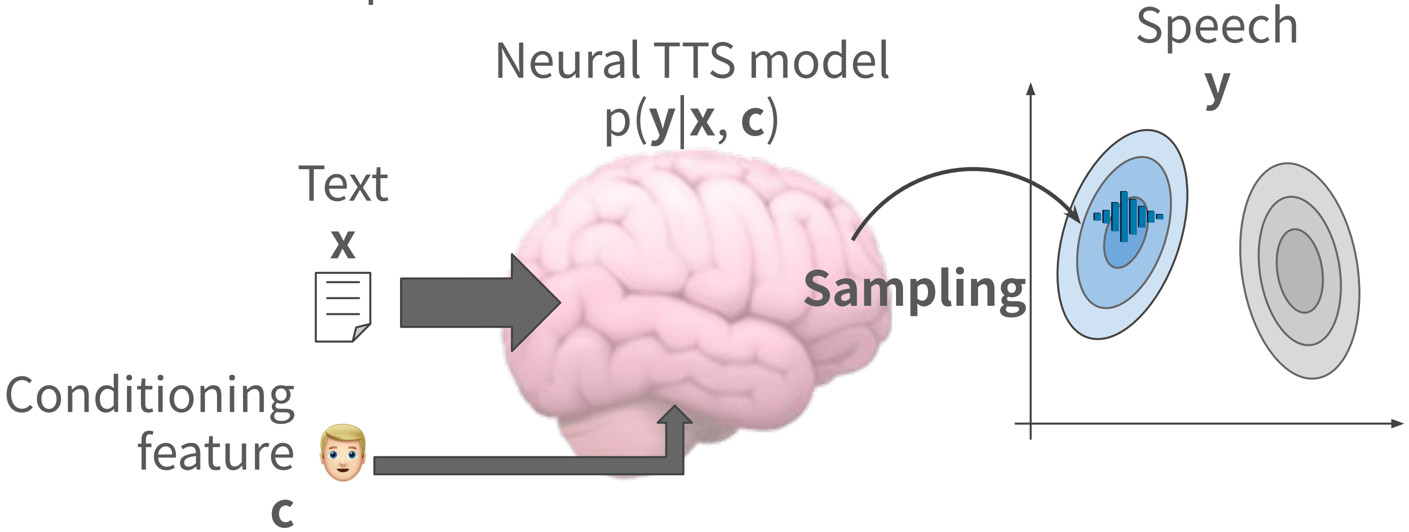

Yuki Saito

Shinnosuke Takamichi

Wataru Nakata

Probabilistic formulation of neural TTS (recap)

- Learning the probability distribution of speech conditioned on text
  - The distribution is **multimodal**, because one text can be spoken from any speakers with any speaking styles → **one-to-many mapping**!
  - To control the sampling process, we can also consider **conditioning feature** as the input to the TTS model.

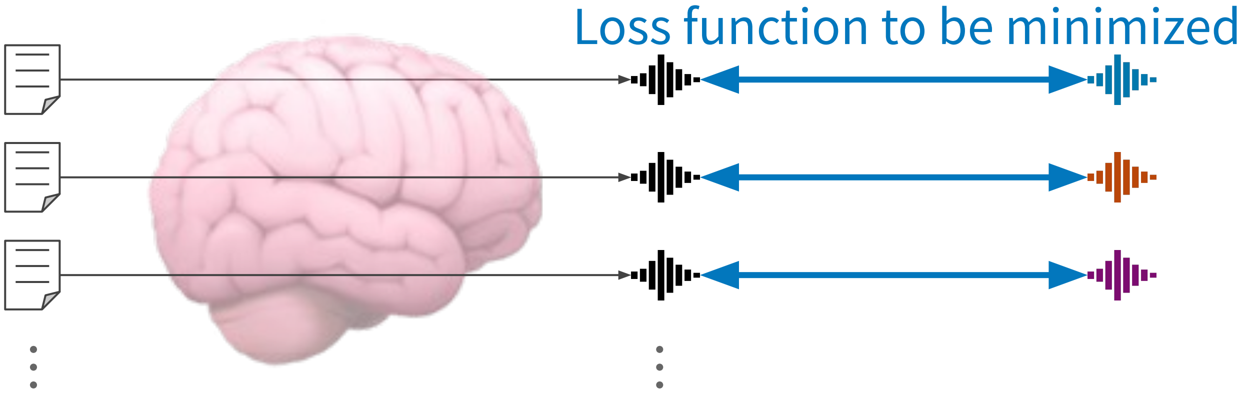


Learning-related topics covered by our tutorial

- Learning TTS from unpaired data
  - **S**elf-**S**upervised **L**earning (**SSL**), multilingual training
- Modeling speech as token sequence
  - Speech tokenization, **L**anguage **M**odeling (**LM**)-based TTS
- Controlling TTS to generate speech that people desire
  - Prompt-based control, preference-aware learning

One challenge in TTS: collecting many text-speech pairs

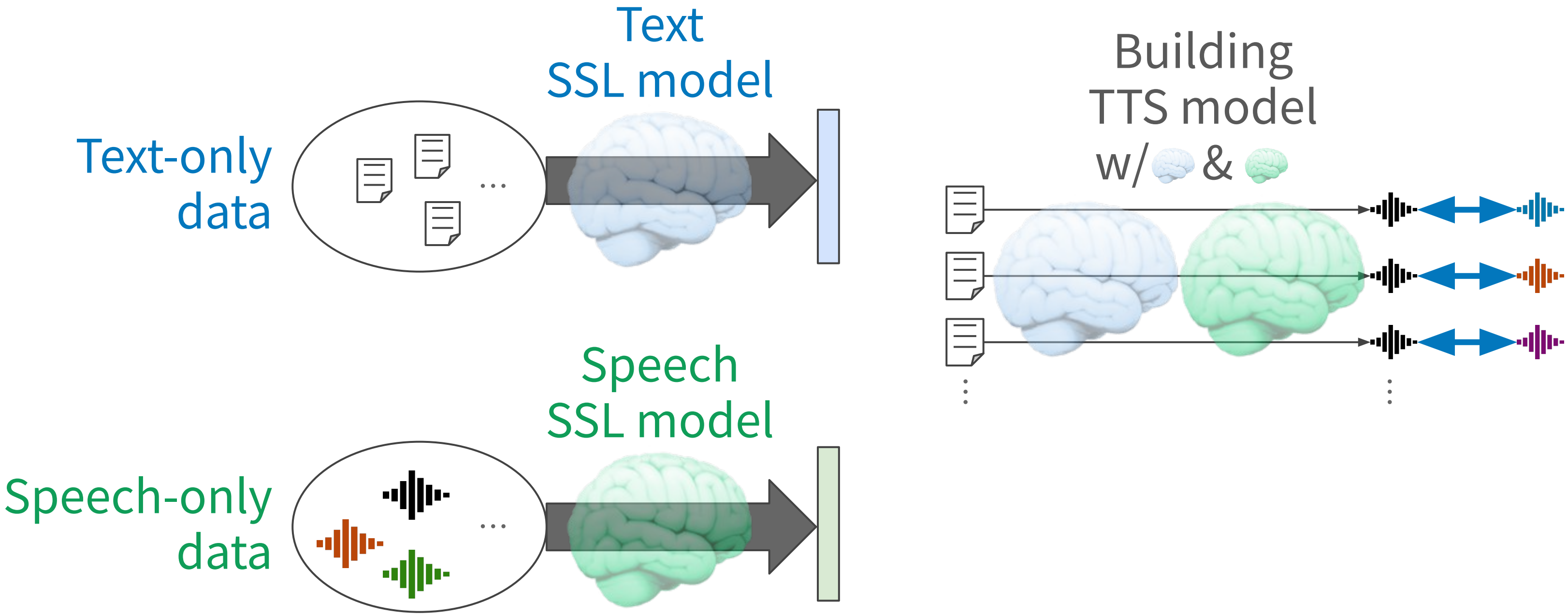
- TTS = seq.-to-seq. & cross-modal mapping from text to speech
  - Paired (text, speech) data are required as the supervision



- Collecting paired (text, speech) data → difficult & even impossible
  - Obtaining text from speech → manual annotation or ASR
  - Obtaining speech from text → speech recording w/ humans

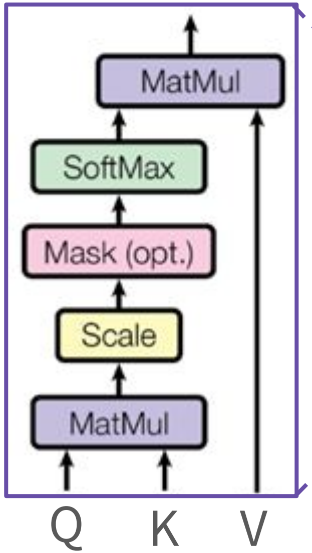
Core technique to cope with this challenge: **S**elf-**S**upervised **L**earning using single-modal data

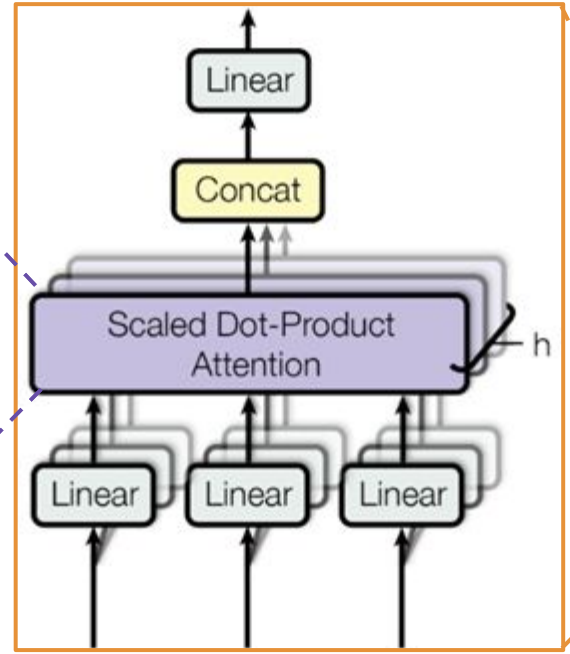
- SSL: learning w/ unsupervised pretext task & supervised target task
  - Pretext task: training SSL model to extract rich feature representations
  - Target task: fine-tuning the features (or model) on specific task



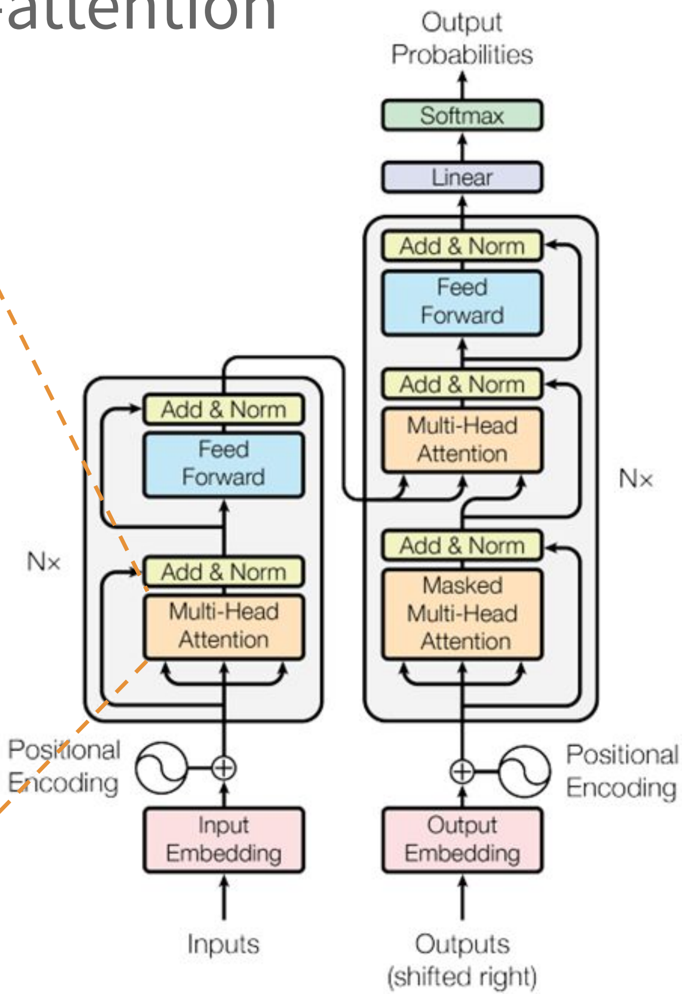
Core architecture: Transformer & self-attention

- Proposed in "Attention is all we need"





One input/output embedding

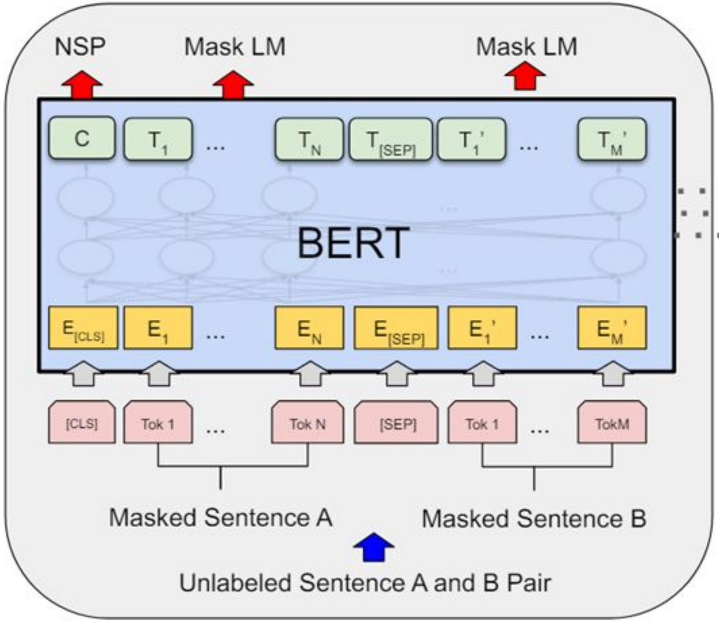


Text SSL example: Bidirectional Encoder Representation from Transformers

- Pretext task: **N**ext **S**entence **P**rediction + **M**ask **L**anguage **M**odel

Predicting masked parts of input tokens from surrounding context/semantic information

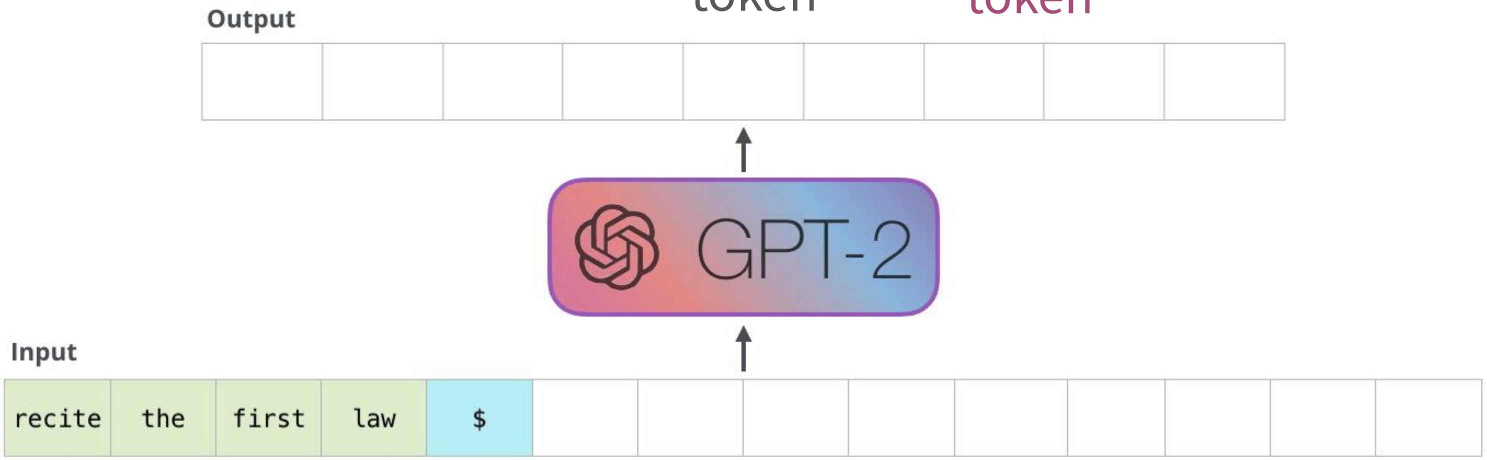
Classifying whether given two sentences are consecutive or not



Text SSL example: causal LM (e.g., GPT pretraining)

- Pretext task: likelihood maximization of **A**uto**R**egressive LM

$$p(x) = \prod_{i=1}^{n} p(\underbrace{s_n}_{\text{Current token}} | \underbrace{s_1, \ldots, s_{n-1}}_{\text{Past token}})$$

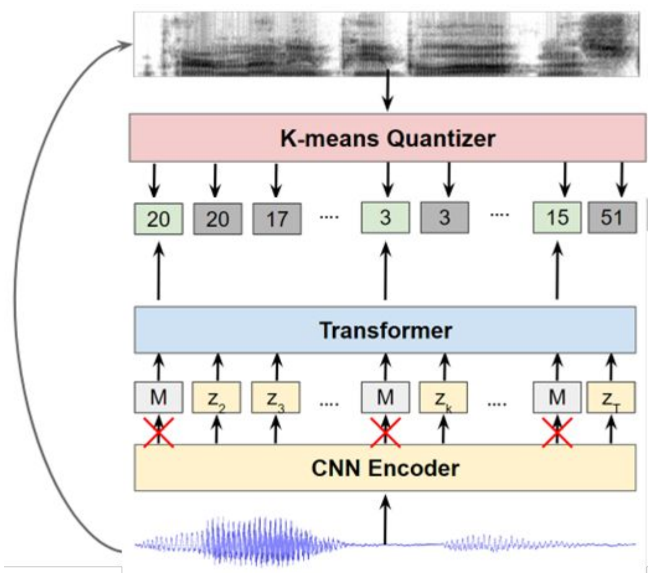


This gif data was taken from [this page](#).

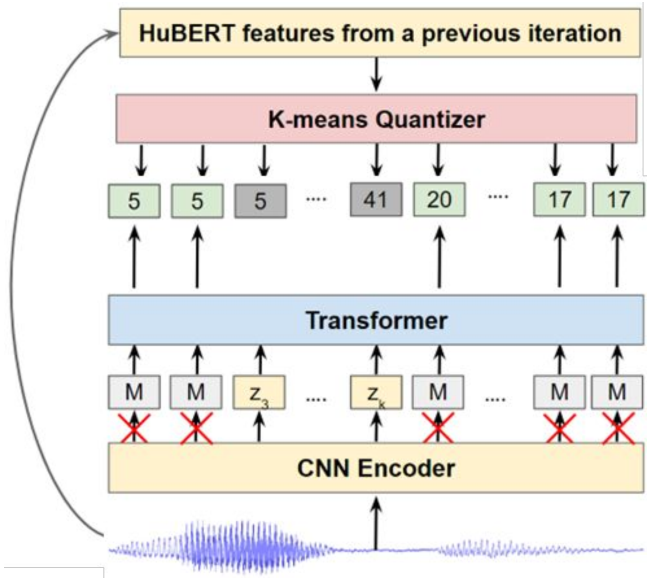
Speech SSL example: Hidden unit BERT

- Two-stage pretext tasks based on MLM on speech representation

Stage 1



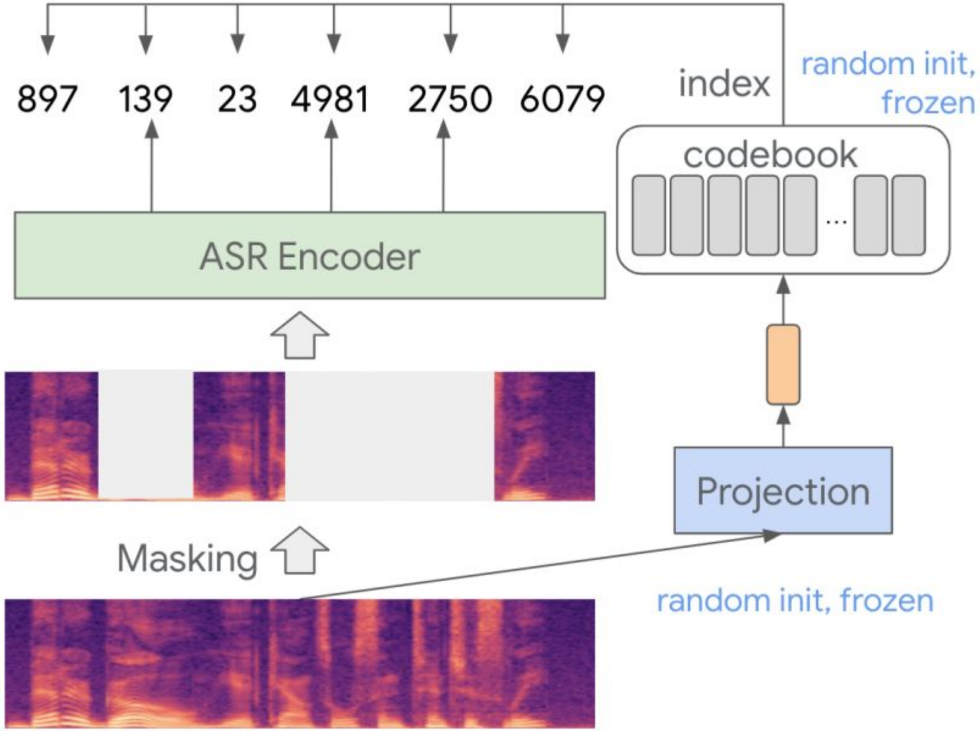
Stage 2



- Improved version: WavLM
  - Considering denoising, speech separation, etc. → improved robustness

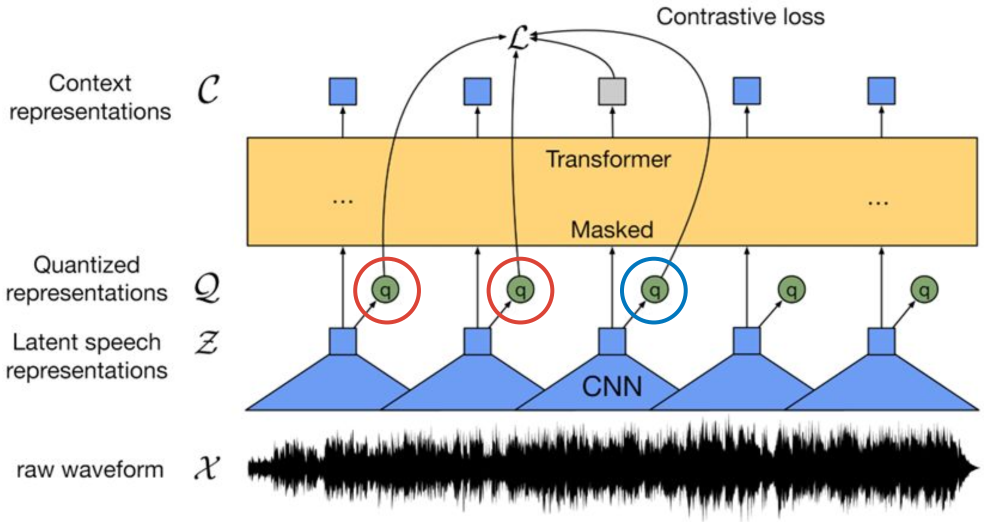
Speech SSL example: **R**andom-**P**rojection **Q**uantization

- Pretext tasks: MLM on randomly initialized & fixed VQ codebook IDs
  - [**BE**RT-based **S**peech pre-**T**raining with **R**andom-projection **Q**uantizer]()



Speech SSL example: wav2vec 2.0

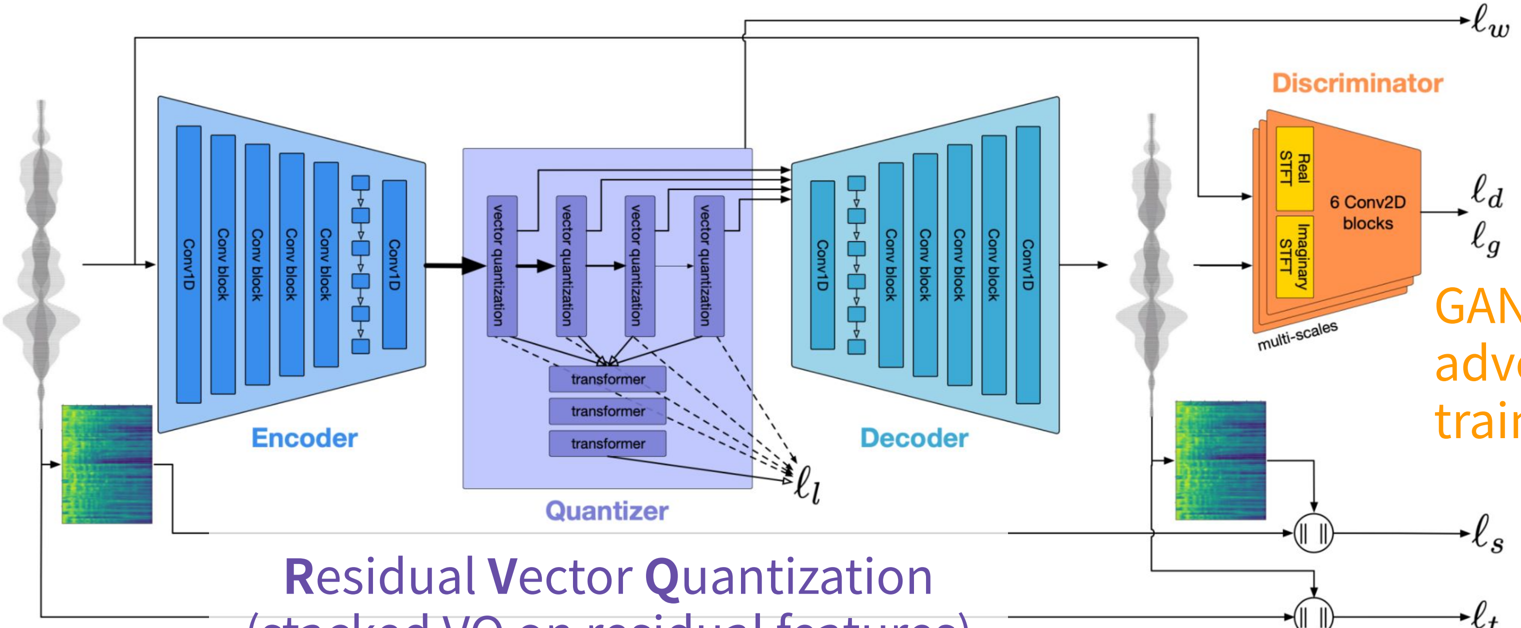
- Pretext task: **C**ontrastive **L**earning w/ contextualized features
  - Learning to make ▢ predict ◯ & unpredict ◯



- Improved version: XLS-R
  - Trained on massive multi-lingual datasets

Speech SSL example: **N**eural **A**udio **C**odec

- Pretext task: waveform reconstruction through autoencoding
  - e.g., [EnCodec](): autoencoding + RVQ + GAN

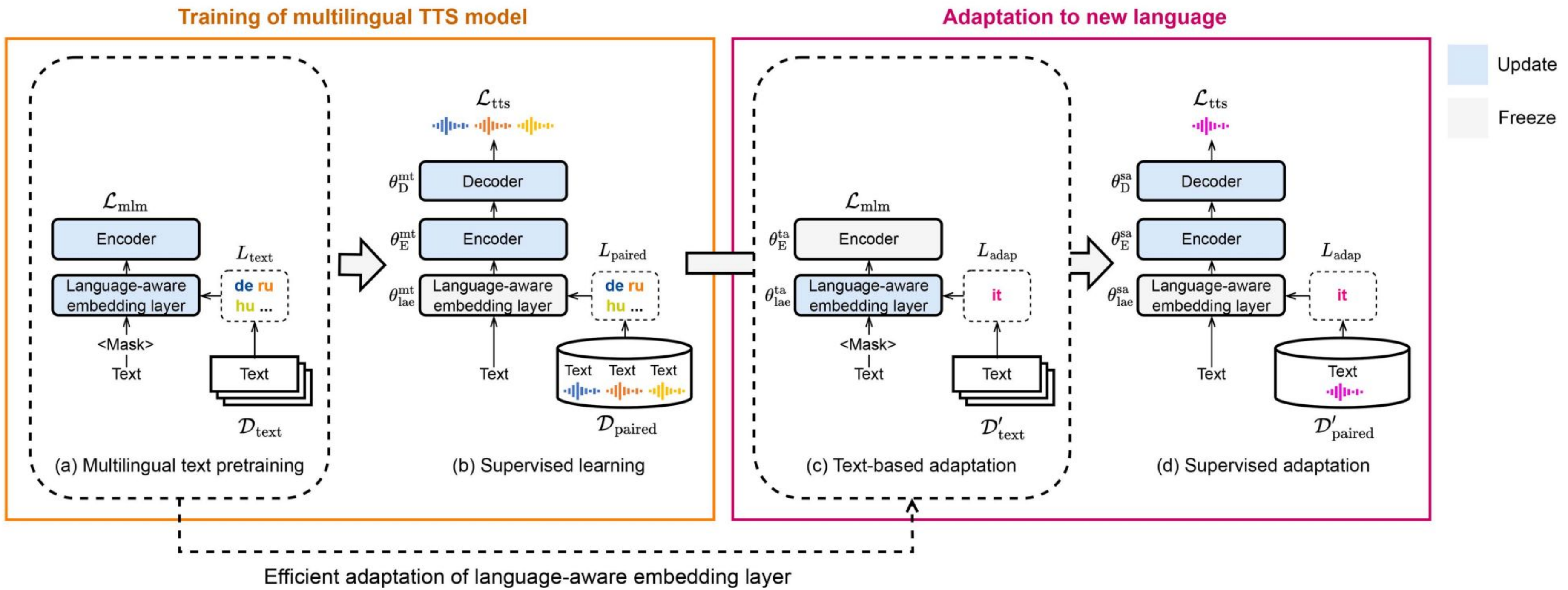


GAN-based adversarial training

Residual **V**ector **Q**uantization
(stacked VQ on residual features)

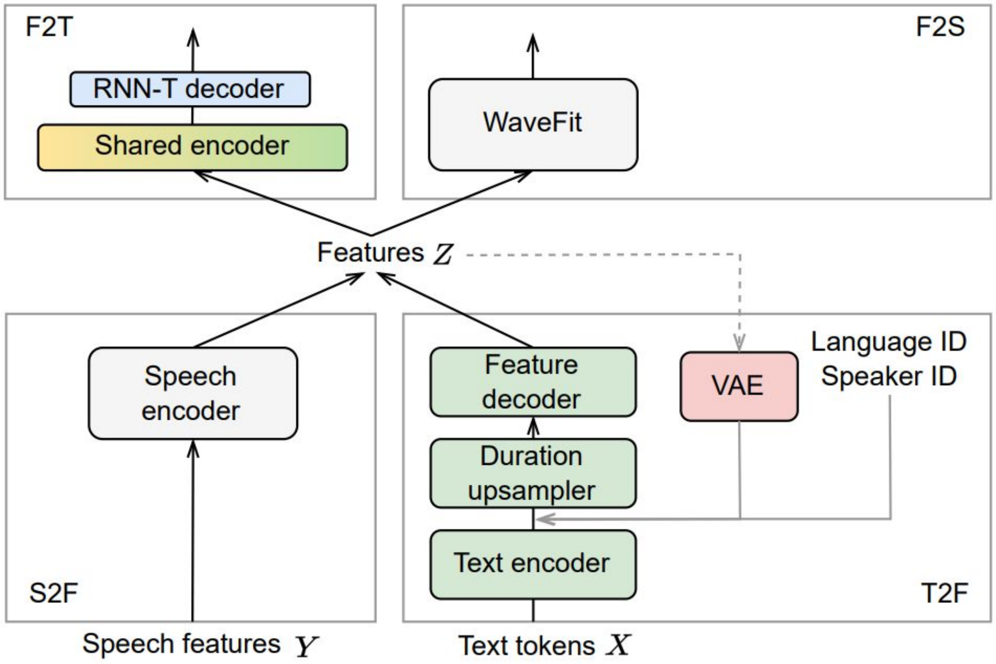
Multilingual TTS w/ SSLed text/speech features

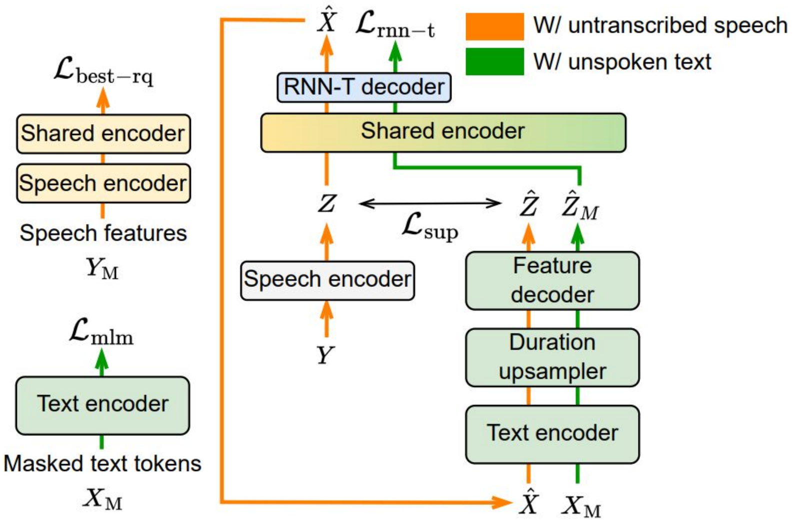
Transfer learning from text-only multilingual pertained model to TTS



Training TTS for 100+ languages without transcribed data

SSL feature as the intermediate representation for ASR and TTS





(a) Pretraining (b) Unsupervised speech-text injection

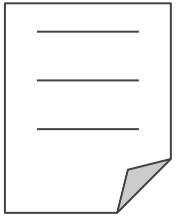
Learning-related topics covered by our tutorial

- Learning TTS from unpaired data
  - **S**elf-**S**upervised **L**earning (**SSL**), multilingual training
- Modeling speech as token sequence
  - Speech tokenization, **L**anguage **M**odeling (**LM**)-based TTS
- Controlling TTS to generate speech that people desire
  - Prompt-based control, preference-aware learning

Differences between text & speech

Information media for communication

- Written language
- Short sequence
  - Length = # tokens
- Only semantic
  - Linguistic information

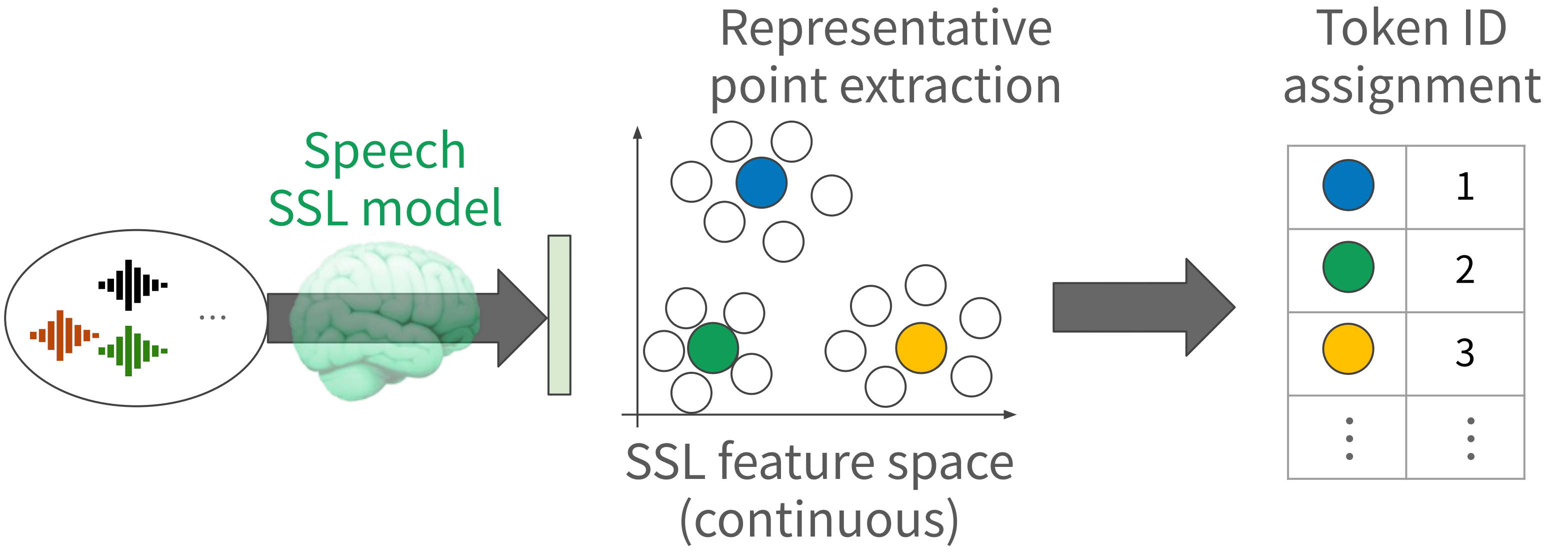


- Spoken language
- Long sequence
  - Length = # samples or # frames
- Both semantic & acoustic
  - Linguistic, para-/non-linguistic (style, speaker, etc.) information

Q: What is effective feature representation of speech for TTS?

Key idea: tokenization of speech SSL features

- Discretizing speech SSL features & defining speech tokens
  - Q: How to discretize speech? → *k*-means or VQ of SSL features

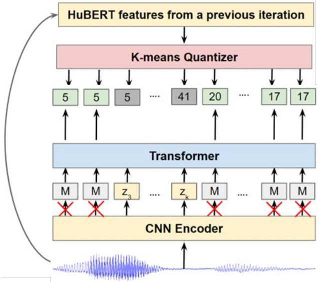


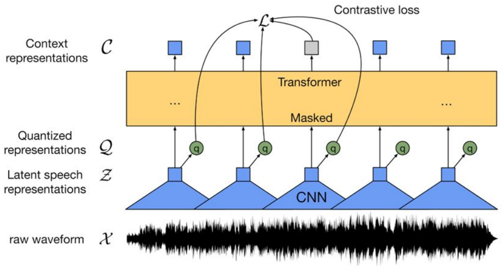
- Representing speech as token sequence like text (char. sequence)
  - **We can introduce knowledge & technique for NLP into TTS!**

Token representation: semantic vs. acoustic

Semantic tokens

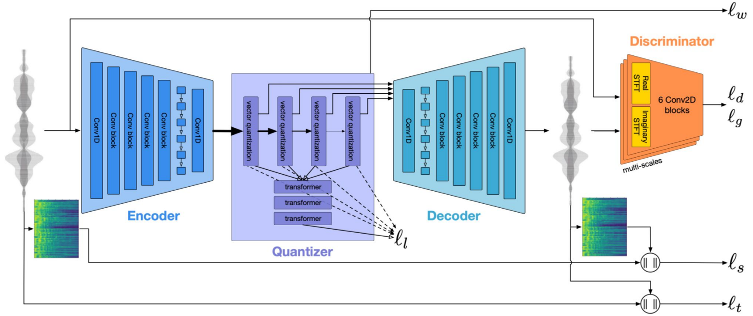
- Trained w/ semantic prediction
  - (Mask) LM, NSP, CL, RPQ
- Suitable for **recognition** tasks
  - Auto. **S**peech **R**ecog.
  - Auto. **S**peaker **V**erification
  - etc.





Acoustic tokens

- Trained w/ reconstruction
  - Autoencoding (w/ RVQ)
- Suitable for **generation** tasks
  - TTS, VC, etc.
  - Can be extended to synthesize general audio

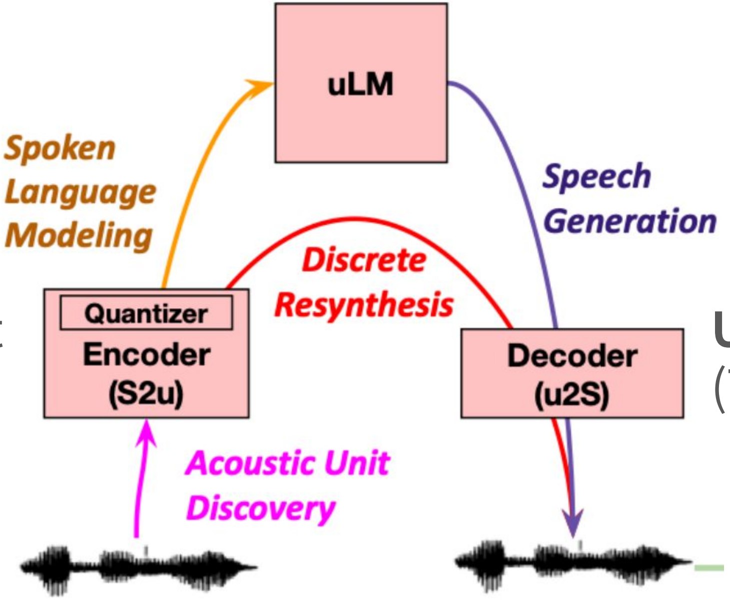


See the tutorial slide by Dr. Yossi Adi to learn these semantic/acoustic tokens.

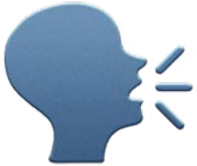
LM for speech tokens (units): [Generative Spoken LM]

Unit LM

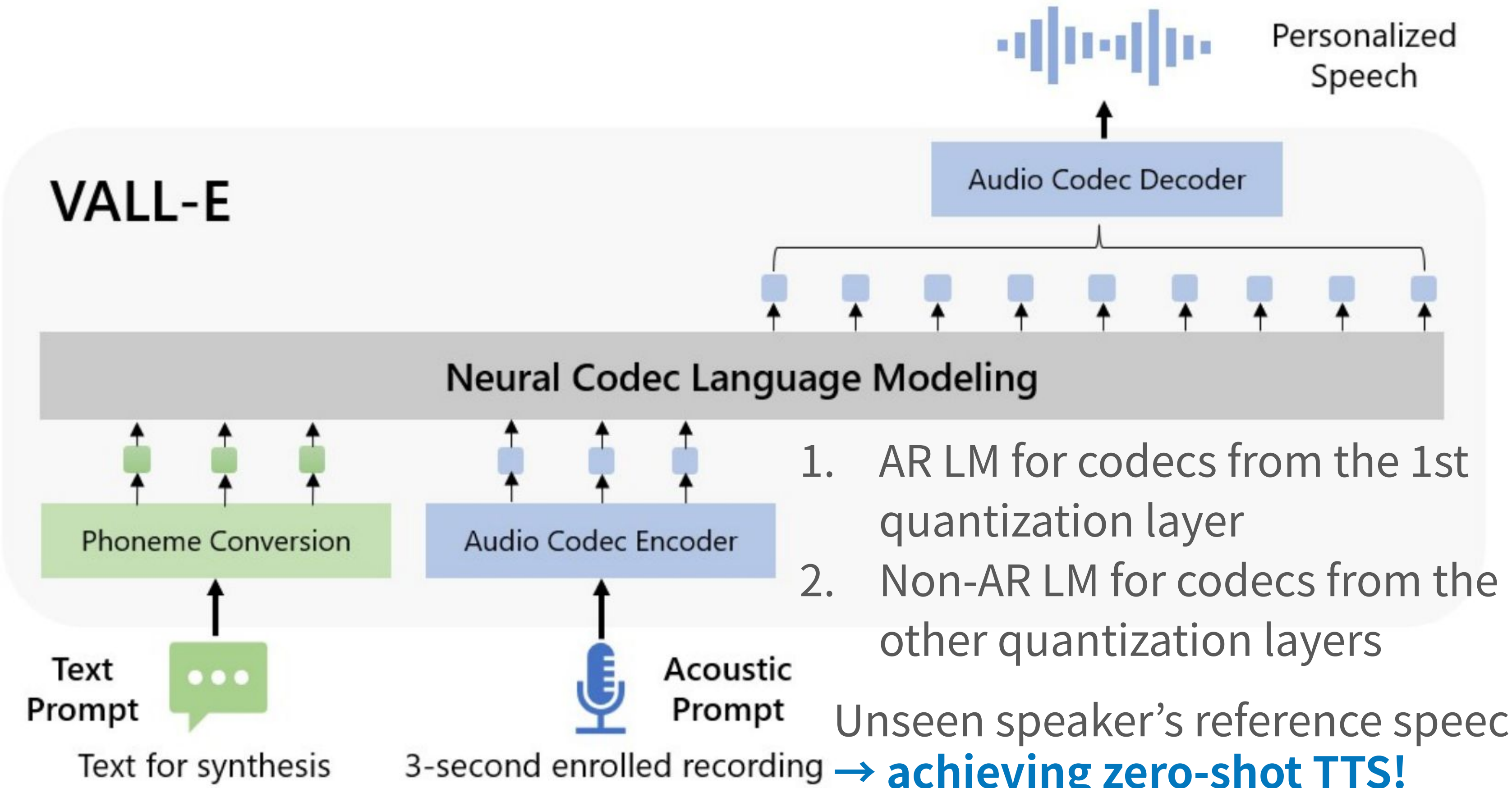
Speech-to-Unit
(HuBERT etc.)



Unit-to-Speech
(TTS + vocoder)



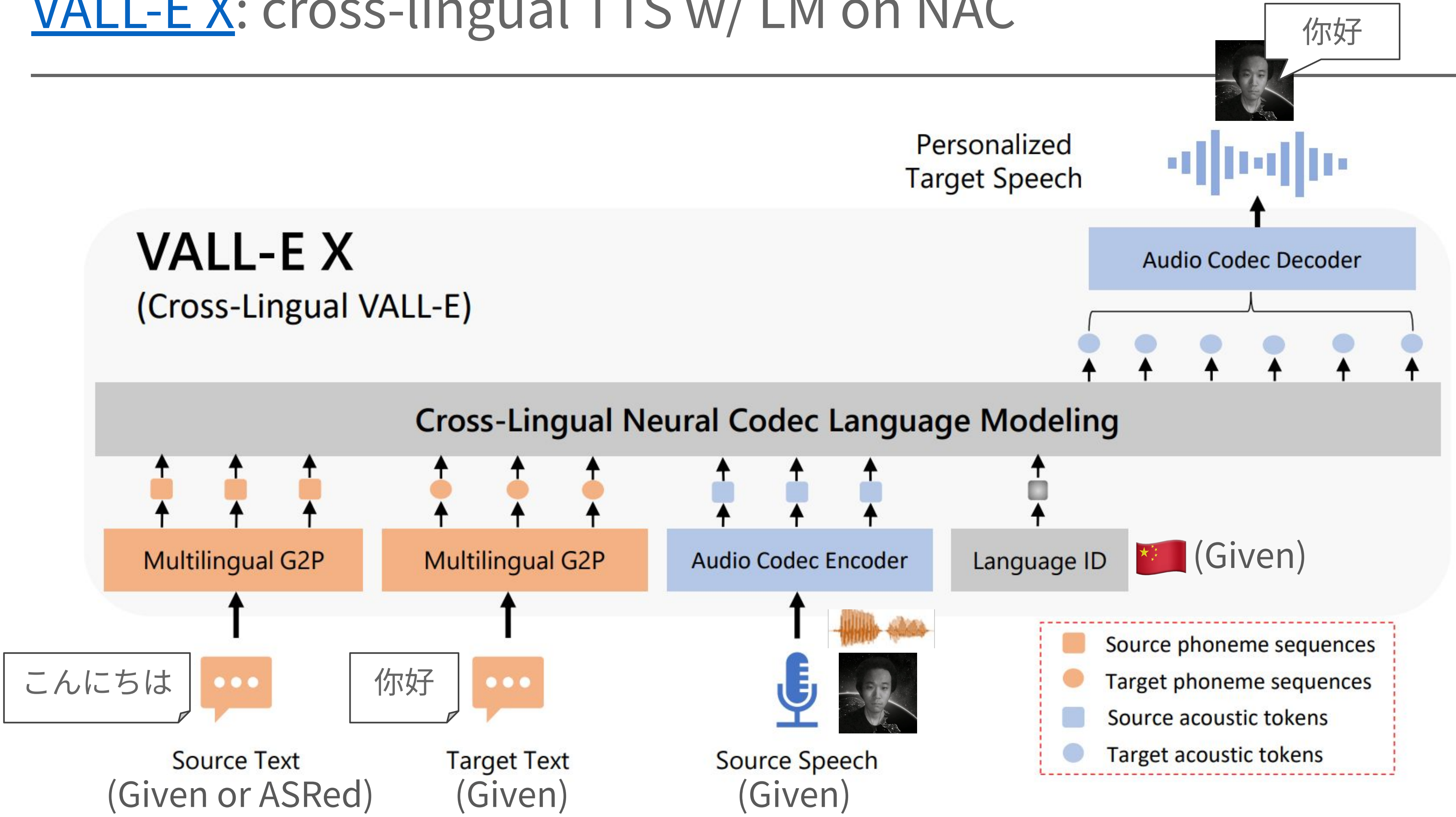
LM-based TTS: LM for acoustic tokens conditioned on input text (e.g., VALL-E)



1. AR LM for codecs from the 1st quantization layer
2. Non-AR LM for codecs from the other quantization layers

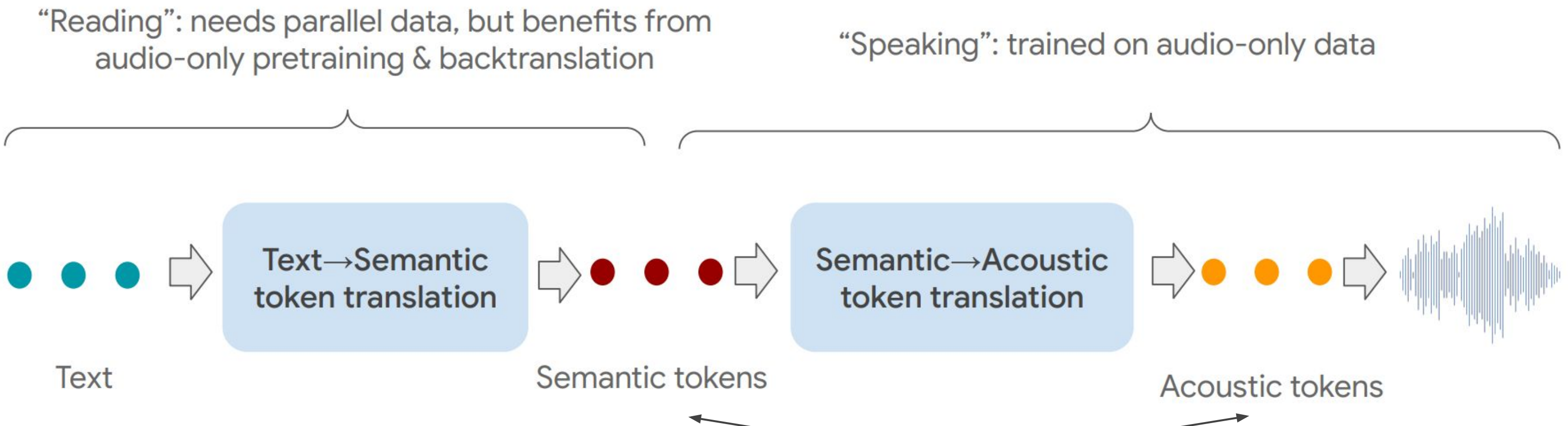
Unseen speaker's reference speech → **achieving zero-shot TTS!**

VALL-E X: cross-lingual TTS w/ LM on NAC



SPEAR-TTS: LM-based TTS with less supervision!

- Split TTS task into two parts: reading & speaking
  - Reading: text-to-SSL feature prediction w/ (easy) supervision
  - Speaking: SSL feature-to-speech prediction w/o any supervision
    - Token conversion ≒ **translation**



Quantized SSL features

Learning-related topics covered by our tutorial

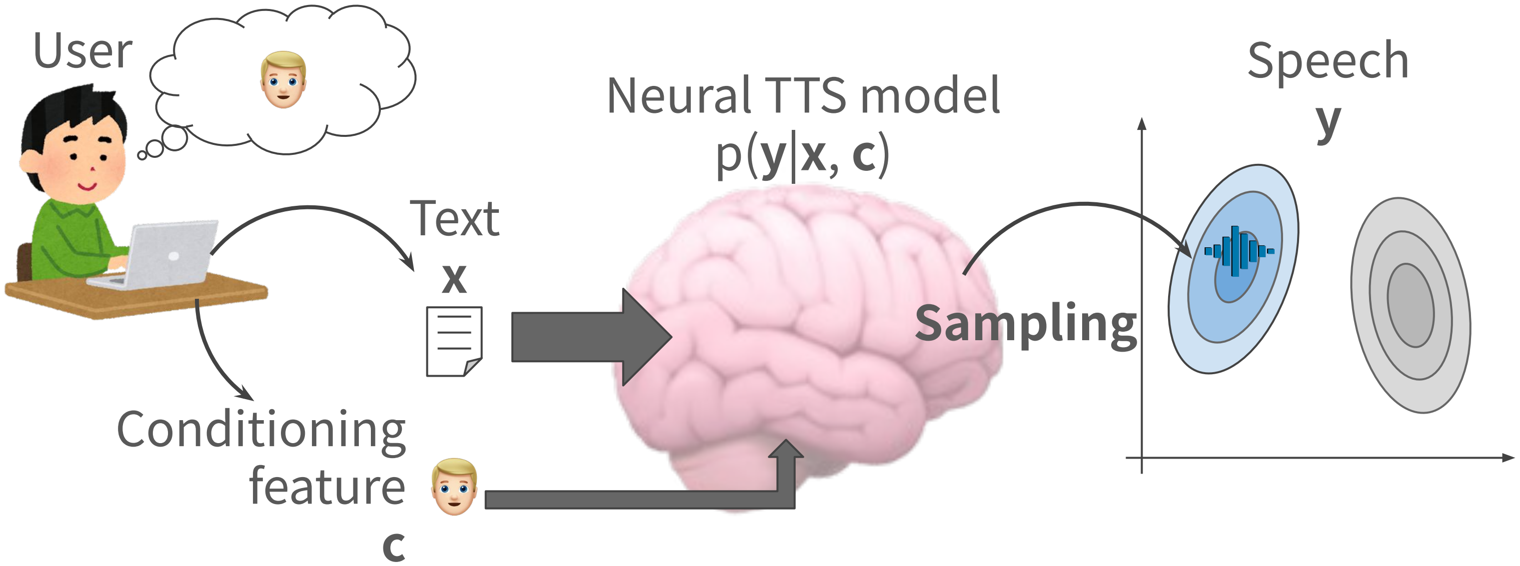
- Learning TTS from unpaired data
    - **S**elf-**S**upervised **L**earning (**SSL**), multilingual training

- Modeling speech as token sequence
    - Speech tokenization, **L**anguage **M**odeling (**LM**)-based TTS

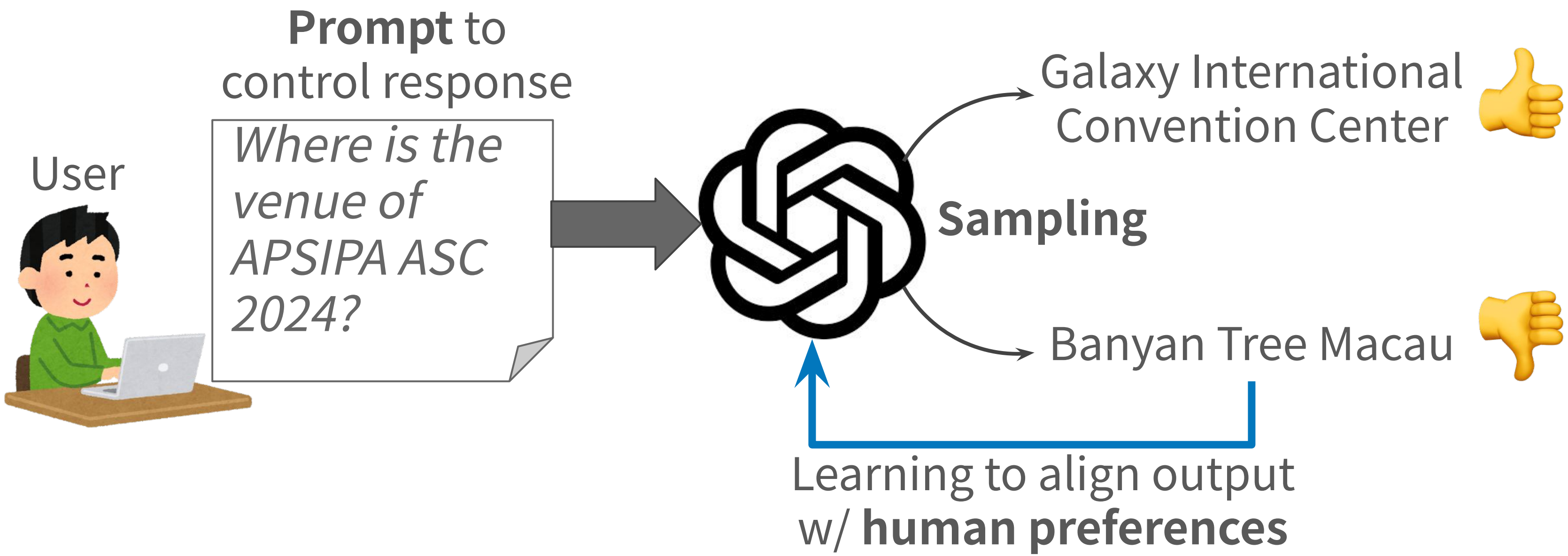
- Controlling TTS to generate speech that people desire
    - Prompt-based control, preference-aware learning

In actual application situation of TTS technologies

- User: input (**x**, **c**) into TTS model to synthesize desired speech
  - Q: **What is the best c that enables user to control synthetic speech?**
    - e.g., reference speech: convenient but not always available
    - The user's desire is often ambiguous & undetermined

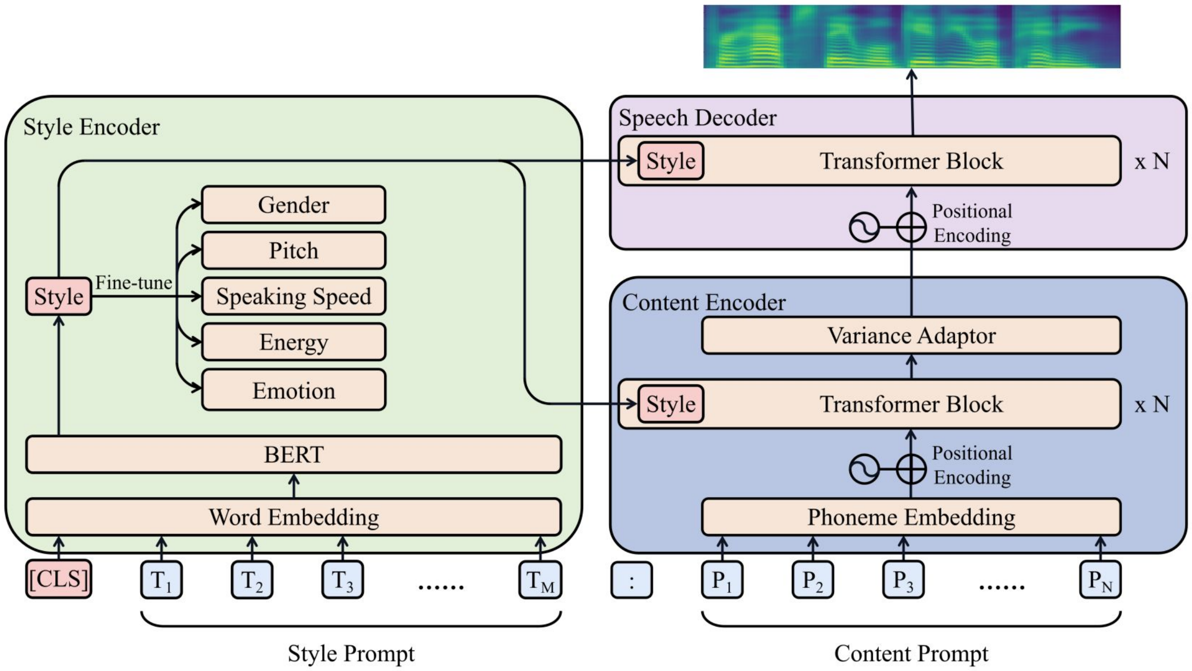


How does ChatGPT work to generate preferable responses?



Q: How can we introduce these techniques to TTS for better control?

PromptTTS: style control of TTS with prompt



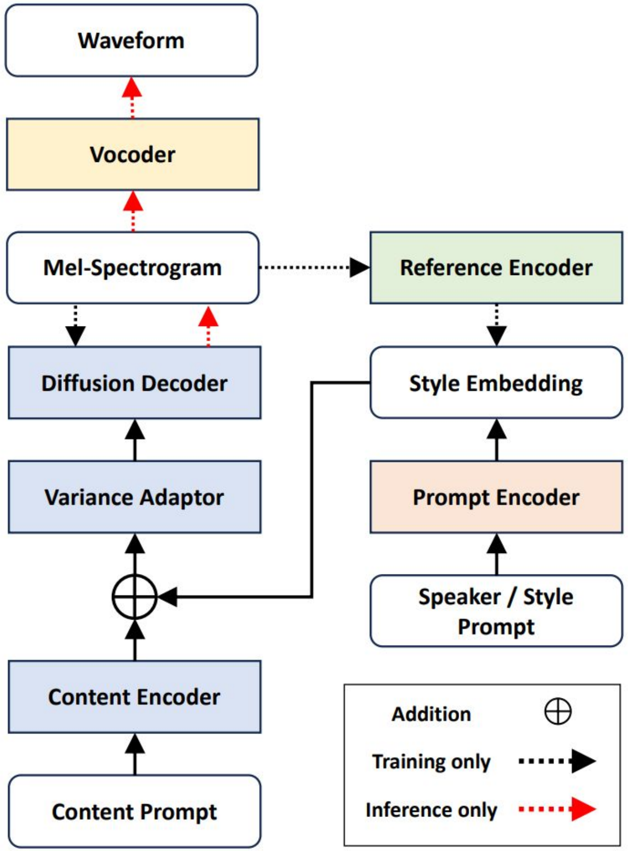
I need a man with a deep and loud voice to talk cheerfully

No important letter come in a parcel, is there?

Style prompt

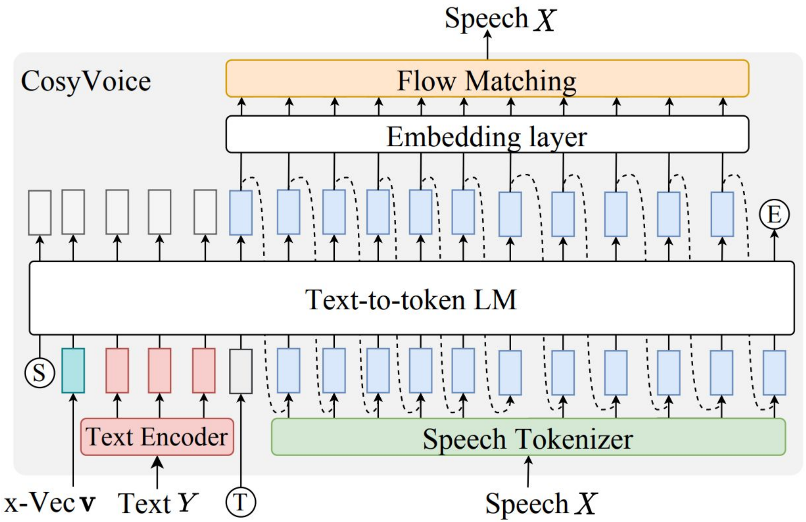
Transcript

PromptTTS++: speaker-identity control of TTS with prompt

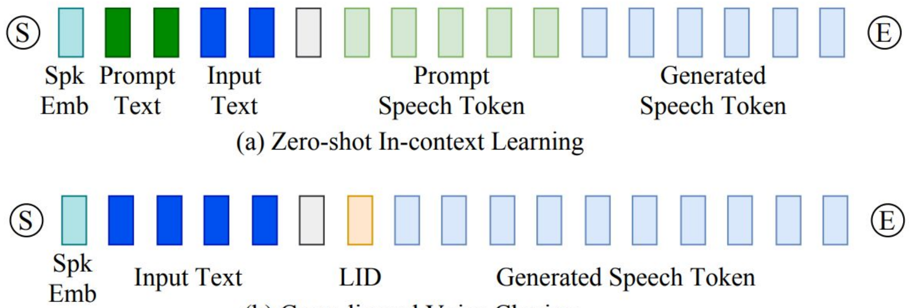


- Free-form style prompt
  - High flexibility but low controllability in speaker identity of speaker
    - e.g., "A woman speaks slowly with low volume and low pitch."
- Categorical speaker prompt
  - More understandable
    - e.g., "The speaker identity is described as soft, adult-like, gender-neutral and slightly muffled."

CosyVoice: multilingual zero-shot LM-based TTS unifying text-/speech-based prompting



Dashed line = inference process



(a) Zero-shot In-context Learning

(b) Cross-lingual Voice Cloning

Speaker Identity
1. Selene 'Moonshade', is a **mysterious**, **elegant dancer** with a connection to the night. Her movements are both **mesmerizing** and **deadly**.<endofprompt>Hope is a good thing.
2. Theo 'Crimson', is a **fiery**, **passionate** rebel leader. Fights with fervor for justice, but struggles with **impulsiveness**.<endofprompt>You don't know about real loss.

Speaking Style
1. A **happy girl** with **high tone** and **quick speech**.<endofprompt>The sun is shining brightly today.
2. A **sad woman** with **normal tone** and **slow speaking speed**.<endofprompt>I failed my important exam.

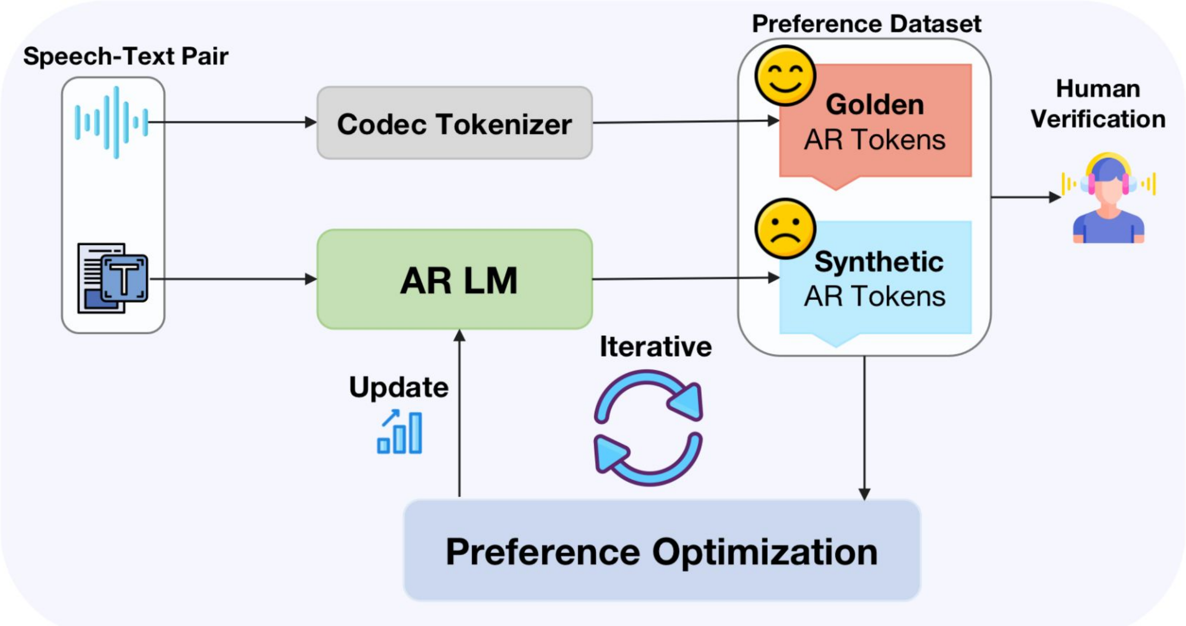
Fine-grained Paralinguistics
1. Well that's kind of scary **[laughter]**.
2. I don't think I over eat yeah **[breath]** and um I do exercise regularly.
3. Well that pretty much covers **<laughter>the subject</laughter>** well thanks for calling me.
4. The team's **unity** and **resilience** helped them win the championship.

Control speaker ID & speaking style w/ natural language description

Synthesize diverse, expressive speech w/ tags

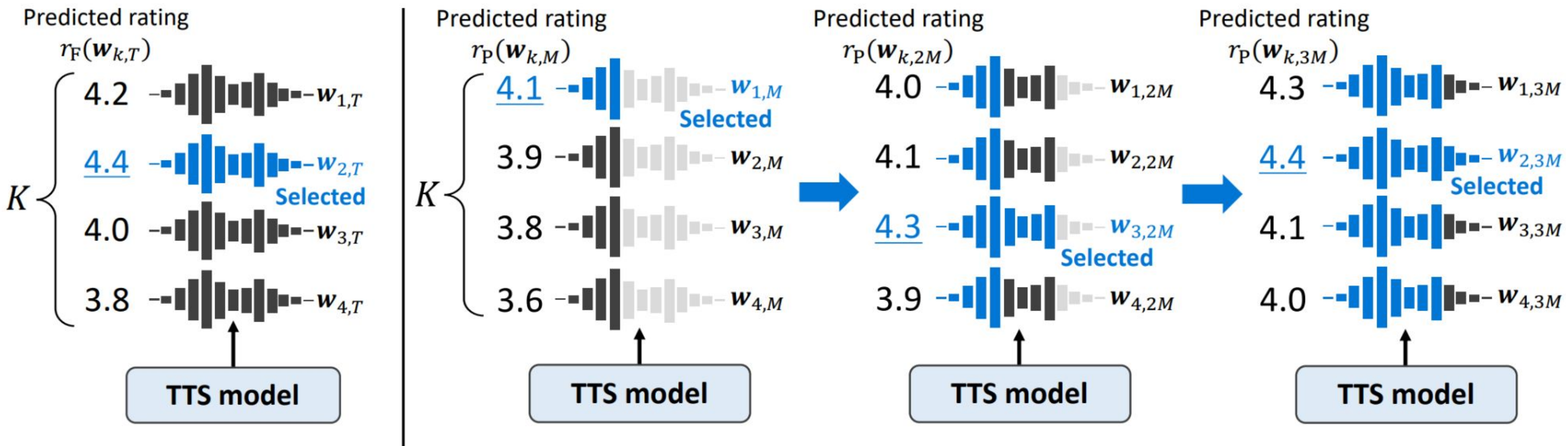
SpeechAlign: iterative optimization of AR LM-based TTS w/ human preferences

Making the distribution of synthetic AR tokens closer to that of golden AR tokens by human preference optimization



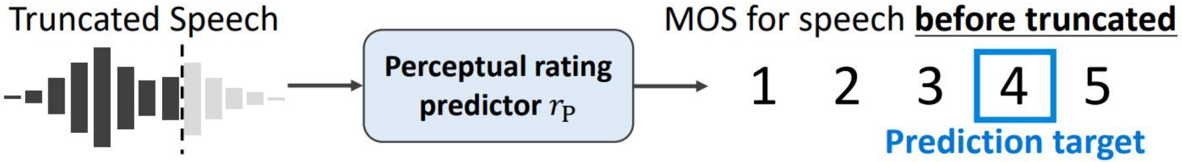
Naturalness-aware TTS sampling: Best-Of-K selection based on Perceptual Rating Prediction

Introducing PRP model to control the token sampling process in TTS



(a) Sequence-wise BOK-PRP

(b) Block-wise BOK-PRP



Q: How can we build such PRP model? → Part 4

Summary of Part 3: learning

- Learning one-to-many mapping from text-to-speech is difficult.
- Using **text/speech SSL** & **LM techniques** is promising research direction to improve the efficiency of TTS model training.
- **Prompting** a trained TTS model offers (probably) an intuitive way to control synthetic speech w/ natural language descriptions.
- Considering **human preferences** for training/inference of TTS model is important for actual scenario, but difficult due to the variability of human subjective evaluation.

Short break (let’s restart at 16:40)

Part 4/5: Evaluation (30 min.)

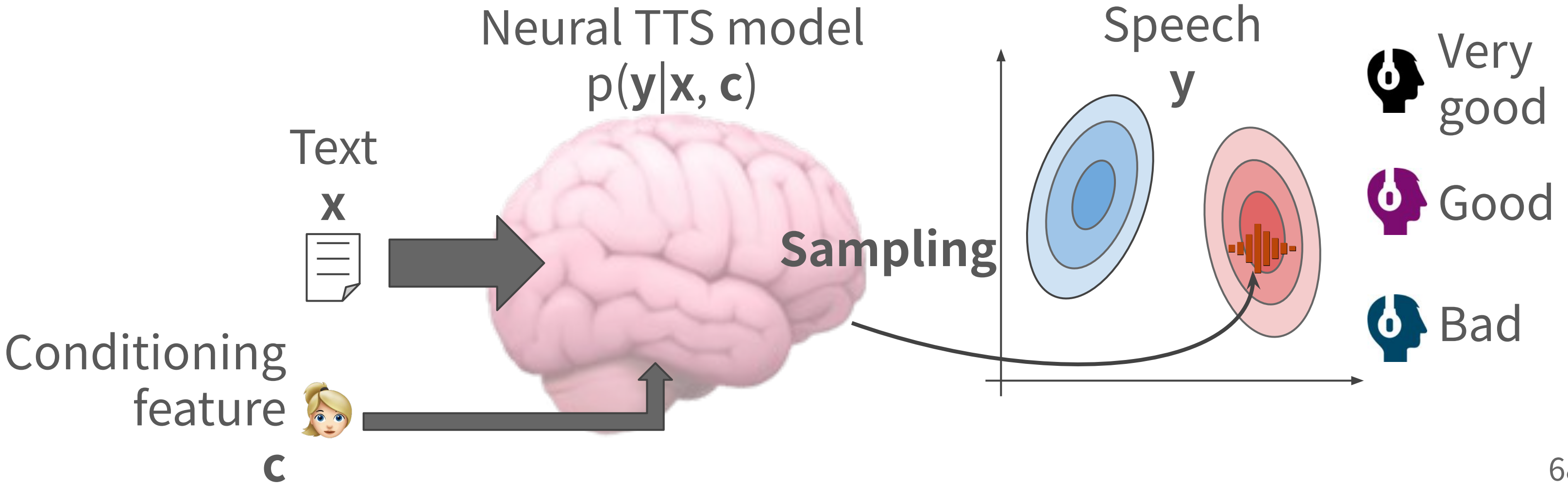
 Yuki Saito

 Shinnosuke Takamichi

 Wataru Nakata

Evaluation of neural TTS (recap)

- Generated speech samples should be **accepted** by human listeners.
  - Human evaluation is intrinsically subjective and difficult to reproduce.
  - Objective evaluation is easy, but not correlated with human evaluation.
  - **Q: How can we universally/fairly evaluate multiple TTS models?**



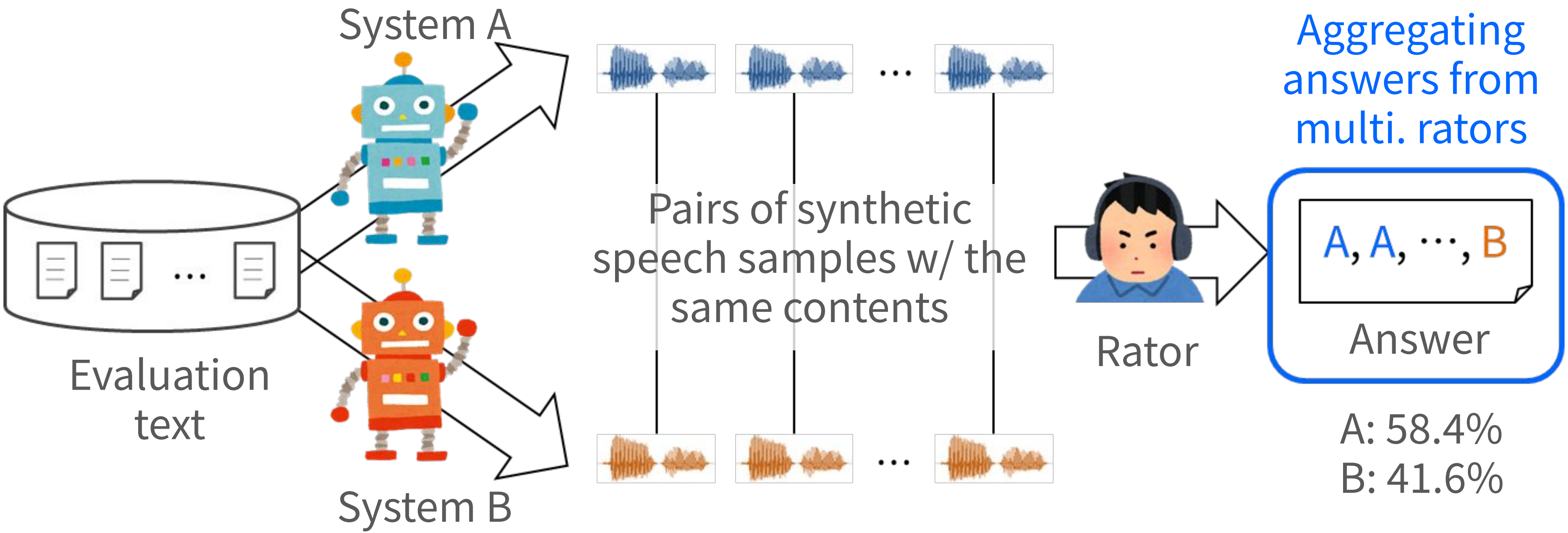
Overview of evaluations for speech synthesis

- Q: What aspect of synthetic speech should be evaluated?
  - A: It depends on the task, but widely used criterias are:
    - **Naturalness**: How human-like does the voice sound?
    - **Speaker similarity**: How well does the voice sound like the target speaker's voice?
    - **Intelligibility**: How clear is the spoken content of the voice?

- Q: Who will evaluate the synthetic speech?
  - A: In many cases, humans will **subjectively** evaluate speech by:
    - Side-by-side (a.k.a.m preference AB) test
    - **M**ean **O**pinion **S**core (**MOS**) test
    - MUSHRA test

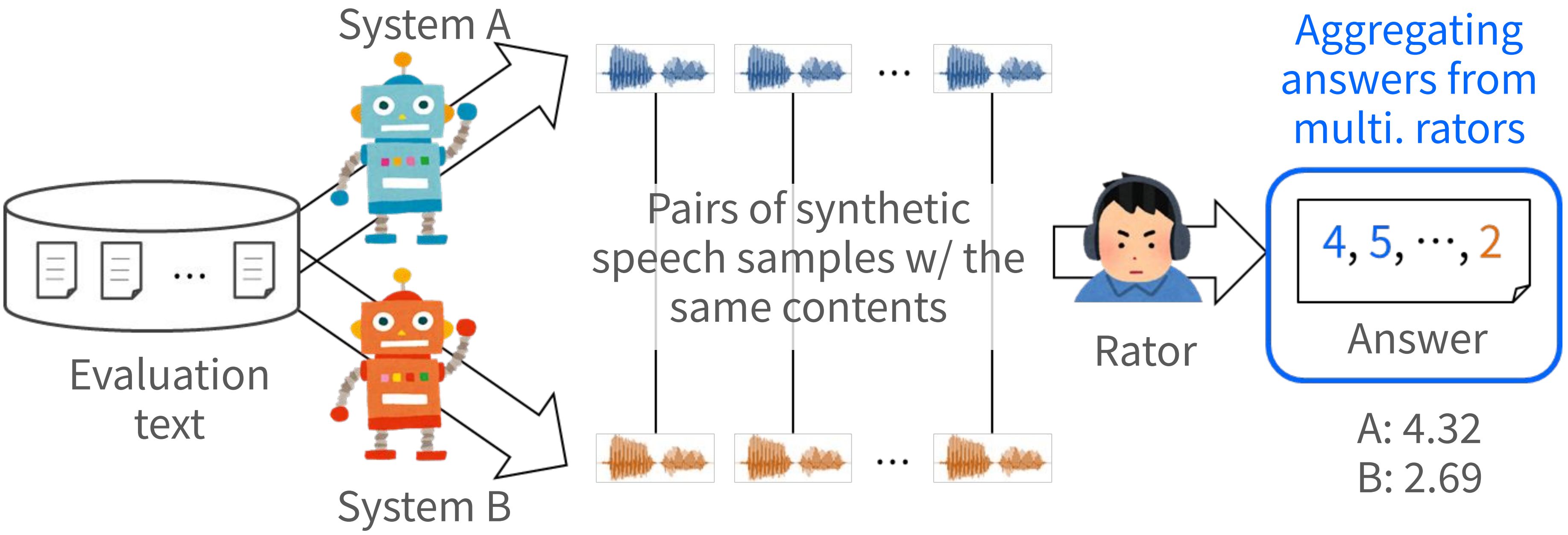
Preference (X)AB test

- Comparing synthetic speech samples by two TTS systems
  - A rater answers which sample sounds better
  - XAB test: using reference speech as X for judging the speech

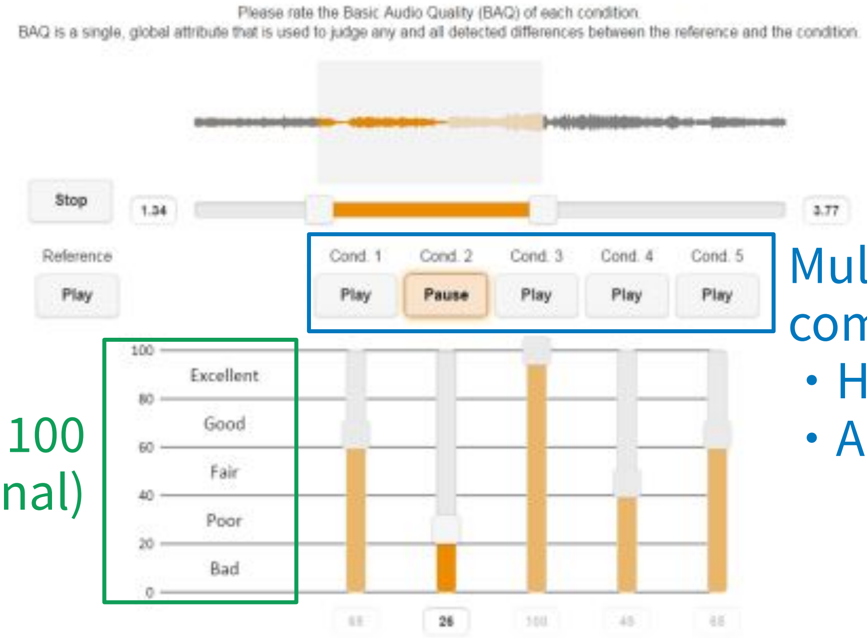


MOS test

- Rating synthetic speech samples with predefined scale
  - Typical scale: 1 (bad), 2 (poor), 3 (fair), (good), 5 (excellent)
  - More than two systems can be compared & evaluated at once



Subjective evaluation example:
MUltiple **S**timuli with **H**idden **R**eference and **A**nchor test



Multi. samples in one comparison including
- Hidden reference: original
- Anchor: heavily distorted

Scale 0 to 100
(grade is optional)

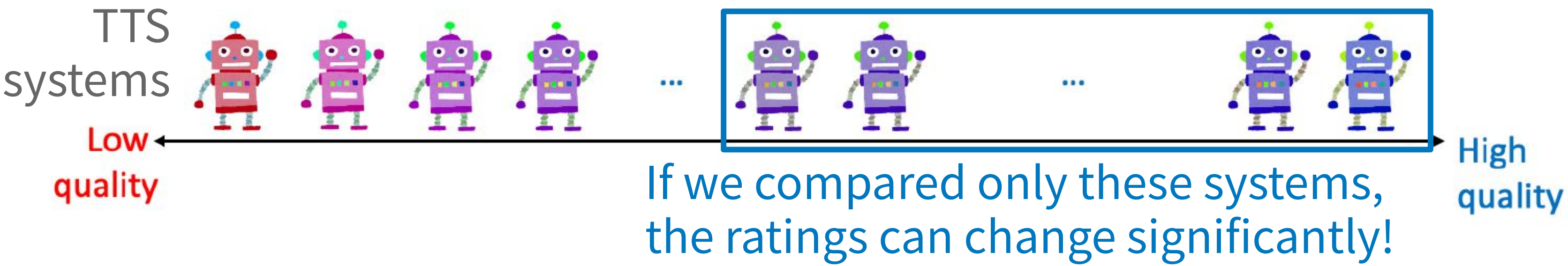
Anchor Reference
(randomized & hidden)

How to design subjective evaluation tests

- Removing evaluation bias as possible
  - Designing certain methods to be intentionally (dis)advantageous
- Examples of good evaluation design w/ lower bias:
  - Selecting **phoneme-balanced text** to be spoken
  - **Balancing # of samples** to be synthesized for each system
  - **Instructing listeners** to wear headphones during evaluation
  - **Shuffling sample ordering** during test
    - i.e., presenting not only (A, B) but also (B, A) in preference AB
- Typical evaluation counts required to discuss statistical significance
  - 150 ~ 200 for each system, but presenting too many samples can make listeners tired 😫

Issues on subjective evaluation

- High cost (i.e., rewards for listeners) to conduct evaluation
  - Poor scalability with respect to # of systems to be compared
- Low reproducibility due to the variability of human judgements
  - Judgement by raters can vary due to many factors (e.g., ordering)
    - Range-equalizing bias: raters in MOS test tend to use the entire range of choices on the rating scale (e.g., from one to five)



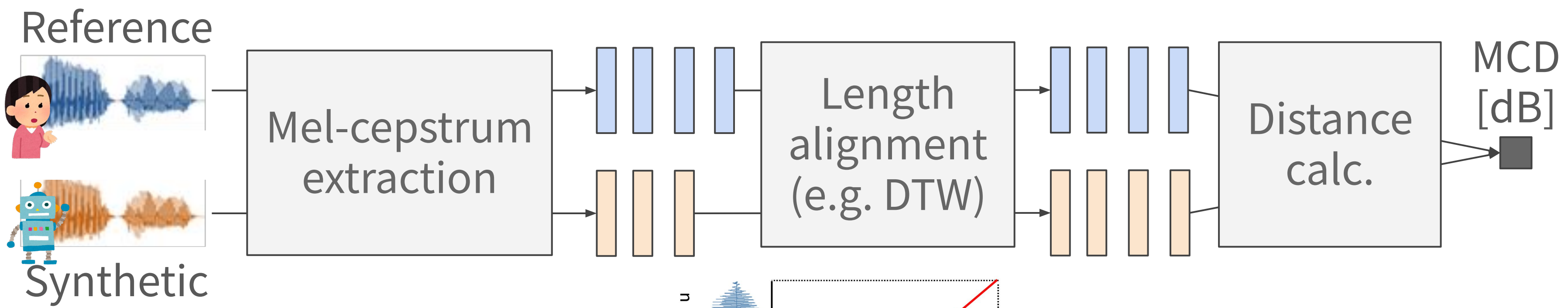
  - At least the metric, # of participants, and # of samples used for the evaluation should be described in paper

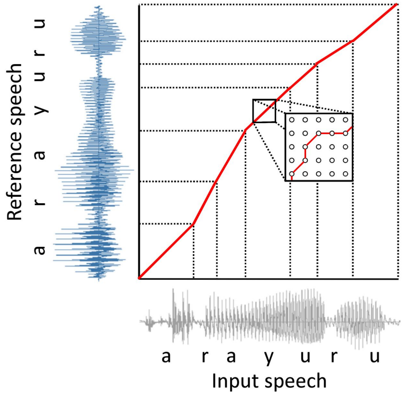
Objective evaluations without requiring listening tests

- Speech-feature-based methods
  - **M**el-**C**epstral **D**istortion, (log)F0 **R**oot **M**ean **S**quared **E**rror, etc.
- Recognition-/Verification-based methods
  - ASR Character Error Rate, **S**peaker **V**erification **A**cceptance **R**ate, etc.
- DNN-based methods
  - Supervised: MOSNet, UTMOS, etc.
  - Unsupervised: SpeechLMScore, SpeechBERTScore, etc.
  - Widely studied in recent years, w/ international competitions on automatic MOS prediction = **V**oice**MOS** **C**hallenges (VMC)

Speech-feature-based objective evaluation

- Extracting speech features and comparing reference w/ synthetic
  - e.g., MCD = distance between spectral features of speech

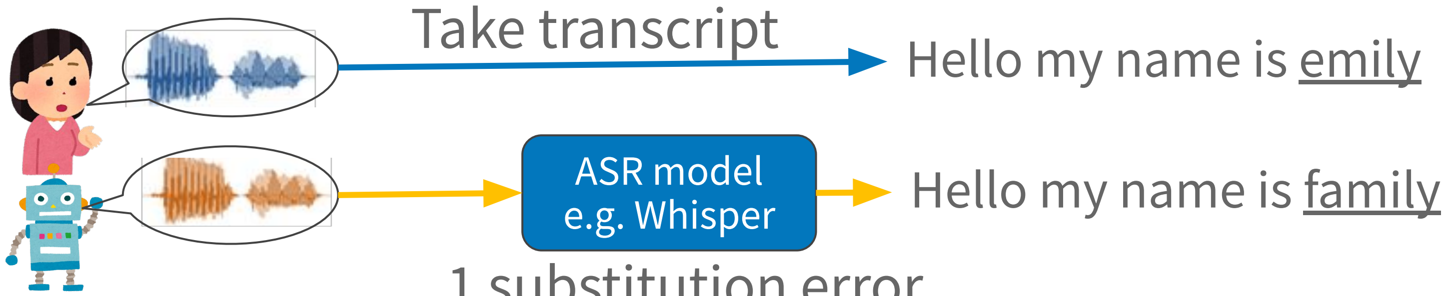




$$\mathrm{MCD}_k = \frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\sum_{k=1}^{K}(c_{t,k}-\hat{c}_{t,k})^2}$$

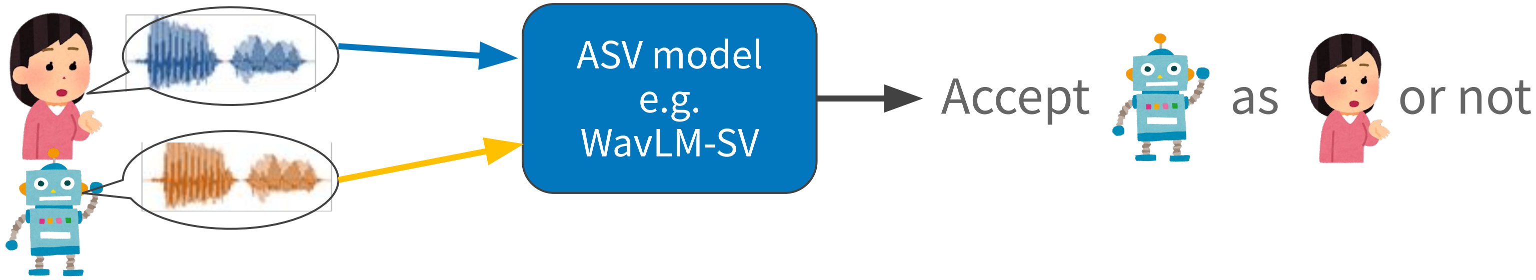
Recognition-/Verification-based objective evaluation

- ASR-based: **W**ord **E**rror **R**ate etc. → intelligibility evaluation



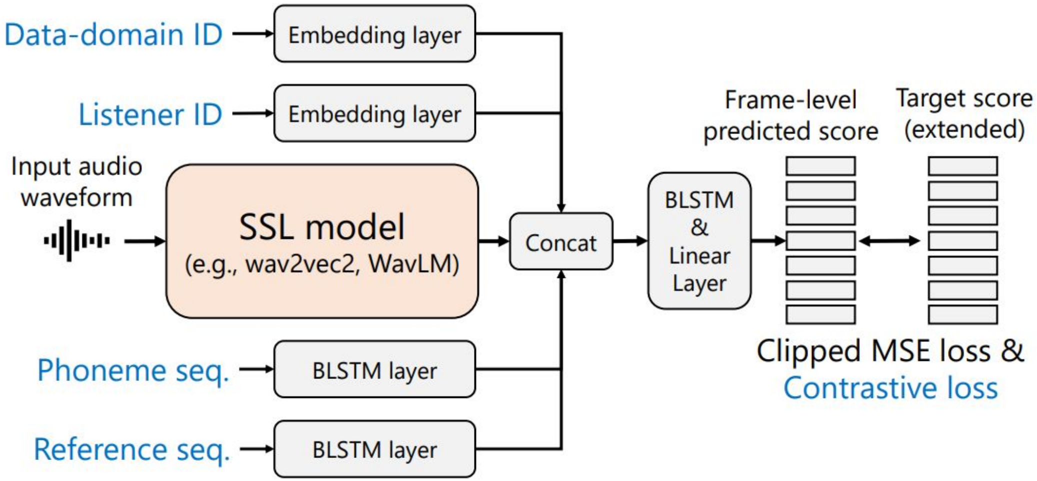
$$\text{WER: } \frac{\text{1 substitution error}}{\text{5 total words}} = 20\%$$

- ASV-based: SVAR etc. → speaker-similarity evaluation



DNN-based objective w/ supervised learning

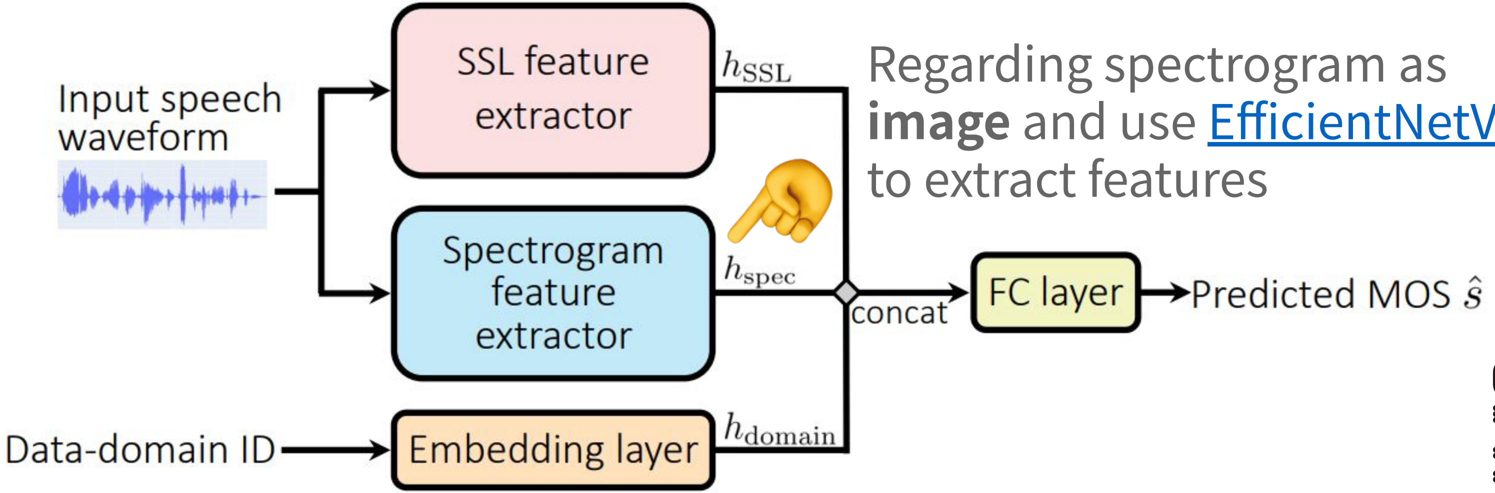
- Building DNNs that predict subjective eval. results of input speech
  - e.g., UTMOS: well-performing system for [VMC 2022]
- UTMOS strong learner: speech SSL model + some conditioning
  - In the actual competition, stacking w/ some weak learners (Ridge, SVMs, etc.) was performed



Codes

UTMOSv2 for VMC 2024 Track 1 (presented in SLT 2024)

- VMC 2024 Track 1: zoomed-in MOS test
  - Re-conducted MOS test w/ better speech synthesis systems only
- UTMOSv2: two feature extractors w/ multi-stage learning
  - **Ranked 1st in 7/16 metrics at the VMC 2024 Track 1** 🏆

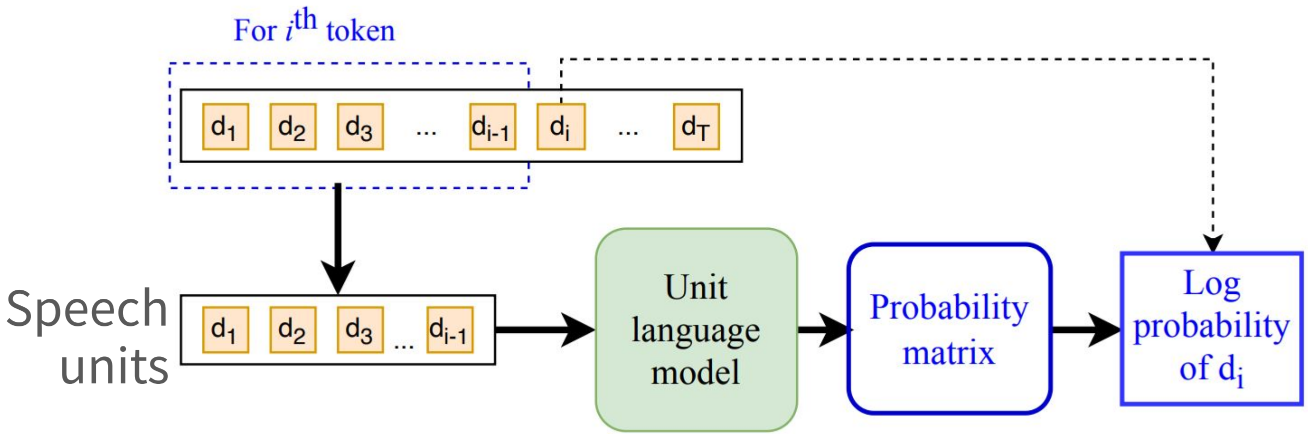


Regarding spectrogram as **image** and use EfficientNetV2 to extract features

Codes

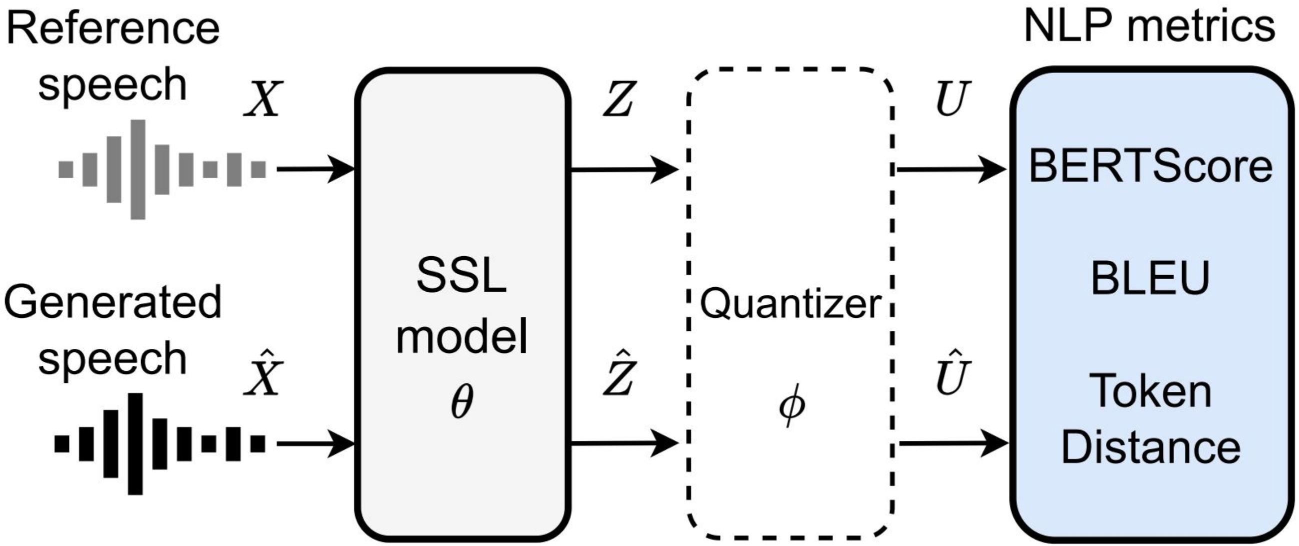
DNN-based objective evaluation w/ unsupervised learning

- Issues: difficulty in collecting training dataset for MOS predictor
  - High cost for annotating MOS to many synthetic speech
- Approach: speech/text SSL for learning w/ unpaired data
  - e.g., SpeechLMScore: computing LM likelihood of speech unit LM
    - Achieving evaluation using synthetic speech only (= **reference free**)



Introducing NLP metrics for DNN-based reference-aware objective evaluation

- Applying NLP metrics for comparing reference/generated speech
  - ∵ **We can tokenize speech by SSL feature quantization!**
  - SpeechBERTScore → similarity as continuous speech features
  - SpeechBLEU, SpeechTokenDistance → similarity as token sequences



How to evaluate objective evaluation metrics

- Q: Can we substitute subjective evaluation with objective ones?
  - What can be the metric for evaluating the feasibility given that we can use subjective evaluation results?
- Metrics used in VMC for performance comparison of MOS predictors
  - **M**ean **S**quared **E**rror: difference between predicted & actual MOS
  - **L**inear **C**orrelation **C**oefficient: widely used correlation measure
  - **S**pearman's **R**ank **CC**: non-parametric, measures ranking order
  - **K**endall's **tau**: measures ranking order, robust to errors
- Evaluation setup: utterance-level or system-level
  - Utterance-level: predictions are made for each utterance
  - System-level: system-wise mean is calculated by the organizer

Rethinking evaluations with the era of foundational speech models

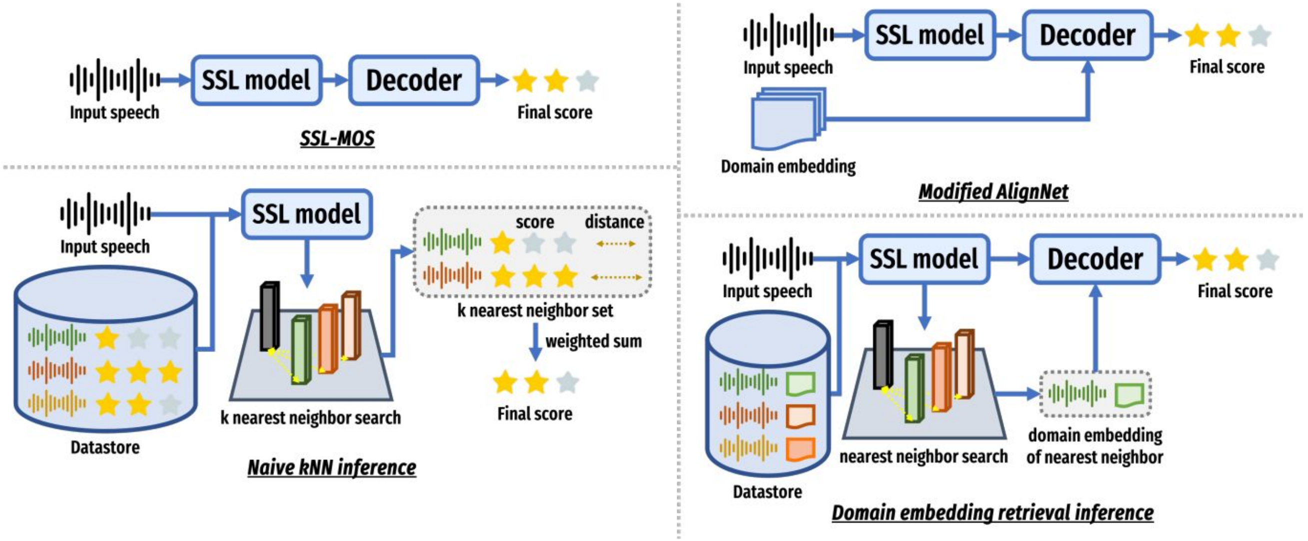
- Beyond naturalness → more complicated criteria, e.g.,
  - Joint evaluation of ASR + LLM + TTS as an AI voice agent
  - Evaluation considering **context**
    - Who will rate the speech?, Where will the rater evaluate it?, etc.
    - Collecting listener metadata (gender, age, etc.) is also important
  - Instruction following ability of prompt-based TTS, audio gen AI, etc.

- Evaluation of of ultra low-resource languages
  - Such speech samples and suitable evaluators are hard to collect

- Necessity for clearly defined & standardized test set
  - The performance of TTS systems can change significantly if different test set (i.e., text) is used for inference.

Benchmark & tools for speech quality evaluation technologies

- Opensourced speech evaluation toolkit and TTS bechmarks
  - c.f., SUPERB for SSL model benchmark, ESPnet
- MOSBench: recently developed benchmark for MOS prediction
  - Collection of MOS dataset for training new MOS predictors



Summary of Part 4: evaluation

- Subjective evaluation, such as MOS test, is the gold standard to compare multiple TTS systems, but suffering from high cost.
- Objective evaluation is an alternative way to reduce the cost, but not always well correlated with the subjective evaluation results.
- Automatic evaluation of subjective quality is a promising approach.
- With the advancement of speech foundation models, we should consider other evaluation **metrics beyond the naturalness**.

Part 5/5: Summary & Outlook (10 min)

Yuki Saito

Shinnosuke Takamichi

Wataru Nakata

Summary of this tutorial

- Emerging Topics for Speech Synthesis: Versatility and Efficiency
  - Goal: Understanding emerging topics in the speech synthesis field
- Neural TTS: using NNs for learning the TTS mapping
  - Essential factors: data, learning, and evaluation
- Takeaways
  - The amount of data available for TTS increase significantly, thanks to the sophisticated data collection/selection/cleansing techniques.
  - SSL for text/speech data can achieve effective training for TTS.
  - Introducing knowledge & techniques of NLP into TTS can cultivate new research direction to develop universal TTS model & evaluation!

Outlook of neural TTS (and its related fields)

- More human-like conversational agents
  - Human like thinking & speaking considering the situation & context
  - Spontaneous & egocentric modeling for human-like speech generation
    - **Q: Do humans speak really natural without making mistakes?**
- Universal speech processor
  - Any modal data can be used to generate speech (any2speech), and speech can create any modal data (speech2any)
- Beyond "speech" synthesis
  - Non-verbal vocalization (e.g., laughing, crying)
  - Non-speech sound (e.g., foley, environmental/instrumental sound)

New research topics related to TTS technologies are waiting for us!

Q&A


Yuki Saito

Shinnosuke Takamichi


Wataru Nakata