# ENHANCING MULTILINGUAL TTS WITH VOICE CONVERSION BASED DATA AUGMENTATION AND POSTERIOR EMBEDDING

Hyun-Wook Yoon, Jin-Seob Kim, Ryuichi Yamamoto, Ryo Terashima, Chan-Ho Song, Jae-Min Kim, Eunwoo Song

NAVER Cloud LY

< Demo >

ICASSP 2024 KOREA

## Motivation

- Creating a multilingual, multi-speaker (MM) text-to-speech (TTS) system is challenging due to the difficulties in collecting polyglot data from multiple speakers.
- To address this issue, we utilize a voice conversion (VC)-based data augmentation method to train the MM-TTS model.
- However, simply including the augmented dataset with the recorded dataset can cause a quality degradation issue. In our case, we observed muffled sound issue in synthesized audio.
- Therefore, we use posterior embeddings (1) to capture the acoustic dissimilarity between the recorded and augmented datasets and (2) to utilize a posterior embedding derived from only the recorded data when synthesizing audio.

## Voice Conversion for Data Augmentation

- **Model**
  - Many-to-many Scyclone model with pitch augmentation [1]
    - Each of monolingual training corpus is reproduced by adjusting pitch values in several semitone-levels to cover a variety of prosodies from multiple speaker and languages.
- **Dataset**
  - Monolingual internal dataset.
  - Korean, English, Japanese, with a single male and female speaker for each.
  - **Number of utterances per speaker**: 500 utterances
- **Process**
  - The original set of 500 sentences is augmented with voices from five other speakers, resulting in a total of 2,500 augmented data. This augmentation process is repeated for each speaker in the dataset.
  - **Outcome**: Through this augmentation, each speaker's original dataset is expanded by a factor of six (3,000), enhancing the diversity and volume of data available for model training.

## Multilingual, multi-speaker TTS system

- **Unified phoneme representation**
  - We integrate 42 English, 47 Korean, and 50 Japanese phonemes into a unified set consisting of **102 phonemes**.
  - We follow the **International Phonetic Alphabet (IPA)** [2] for merging phonemes from different languages and phonemes with similar pronunciations (e.g. 'm', 'n' in nasal sound) are combined. (details are provided in the Table1)

Table1: Unified phonemes table

| CONSONANTS (PULMONIC) | | Unified symbol | Original IPA symbol: ko | jp | en |
|---|---|---|---|---|---|
| Plosive | Bilabial | p | pʰ ㅍ (파랑) | p パ (パン) | p p (pack) |
| | | b | b ㅂ (바람) | b ば (ばしょ) | b b (back) |
| | Alveolar | t | tʰ ㅌ (타다) | t た (たべる) | t t (time) |
| | | d | d ㄷ (다수) | d ど (どうも) | d d (dog) |
| | Velar | k | kʰ ㅋ (크기) | k く (くる) | k k (kiss) |
| | | g | g ㄱ (가방) | g が (がっこう) | g g (gaggle) |
| Nasal | Bilabial | m | m ㅁ (마을) | m ま (まあ) | m m (much) |
| | Alveolar | n | n ㄴ (나무) | n な (なっとう) | n n (note) |
| Fricative | Labiodental | f | | ɸ ふ (ふく) | f f (fish) |
| | Alveolar | s | s ㅅ (사랑) | s さ (さっそう) | s s (soup) |
| | | z | t̤z ㅈ (자유) | z ざ (ざくろ) | z z (zip) |
| | Alveolo-palatal & Postalveolar | sh | | ɕ し (しき) | ʃ sh (ship) |
| | Glottal | h | h ㅎ (하늘) | h は (はな) | |
| Affricate | Postalveolar | ch | | ʧ ち (ちゃ) | ʧ ch (chair) |
| Trill & Approximant | Labiodental | r | | ɾ ラ (ラーメン) | ɹ r (run) |

- **Posterior encoder**
  - We train the **posterior encoder** [3] to focus on capturing the distributions of recorded and augmented data by providing it with explicit speaker and language information.
  - During inference, the encoder selectively retrieves posterior embeddings from the entire recorded dataset within the training set, averaging these to obtain the final posterior embedding.
  - As illustrated in the Figure1, data clusters in the latent space are distinguishable based on their origin from either recorded or augmented data.

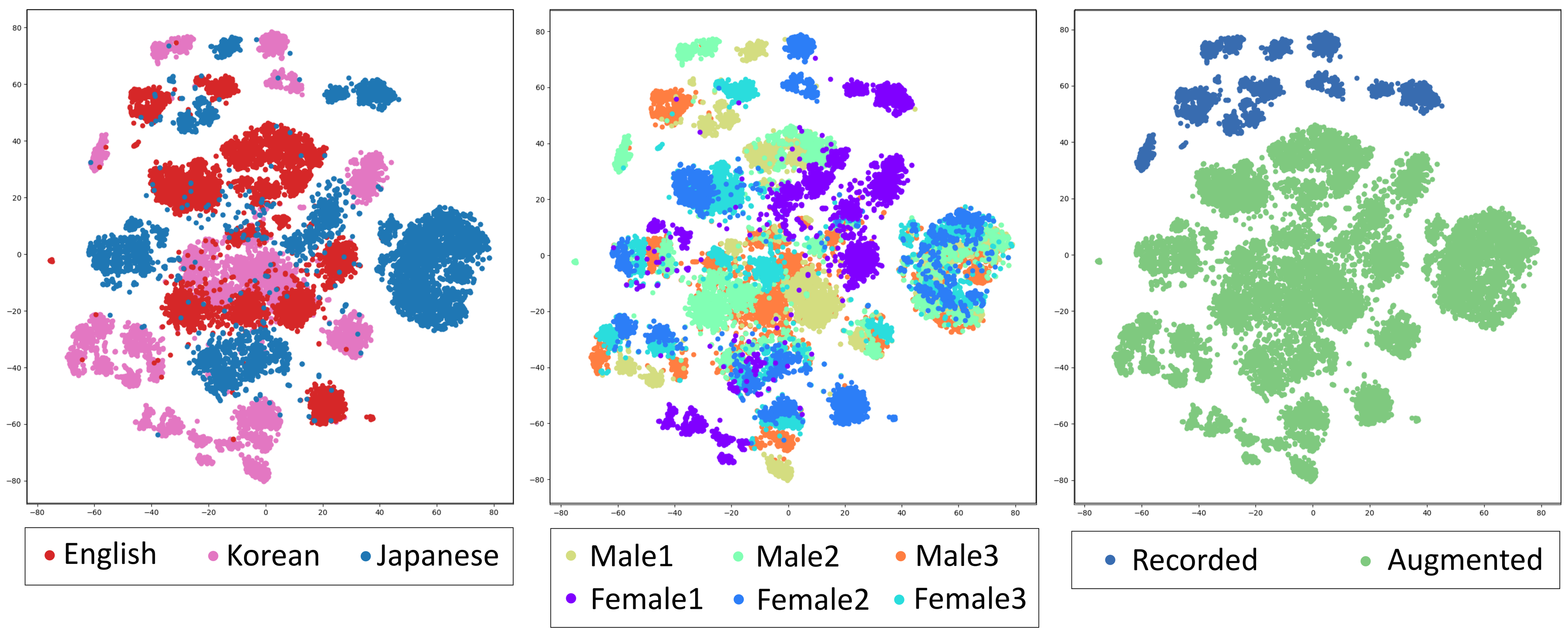


Figure1: t-SNE plots

- **Text-to-speech model**
  - The system includes a **context encoder, a duration predictor, an autoregressive decoder, and a PWG vocoder** [4], complemented by **a speaker and language look-up table** as well as **a posterior encoder**.

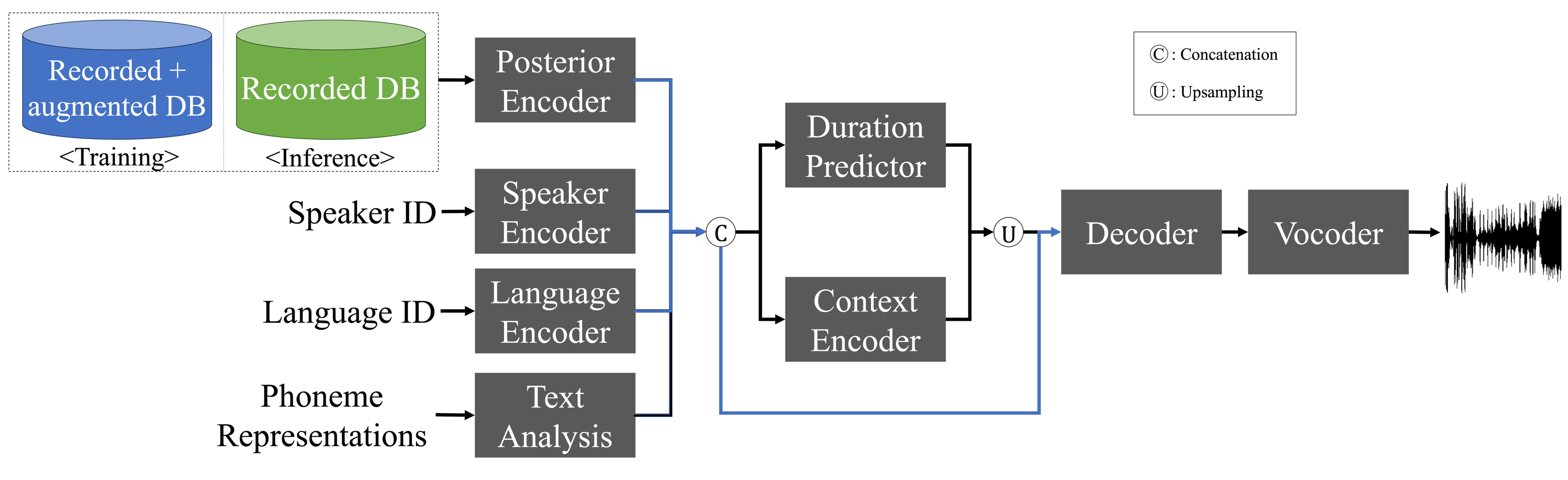


Figure2: Model diagram

## Experiments

- **Compared models**
  - **CM-TTS:** cross-lingual, multi-speaker TTS model
  - **MM-TTS:** VC-augmented multilingual, multi-speaker TTS model
  - **MM-TTS$_{vae}$:** MM-TTS with posterior embeddings
- **Objective evaluation**
  - **Intelligibility:** WER(%), CER(%)
  - **Acoustic similarity:** $F0_{rmse}$(Hz), log spectral distance (LSD)(dB)
- **Subjective evaluation**
  - **Naturalness:** MOS

| Model | English | | | | Korean | | | | Japanese | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WER(%) | CER(%) | $F0_{rmse}$(Hz) | LSD(dB) | WER(%) | CER(%) | $F0_{rmse}$(Hz) | LSD(dB) | WER(%) | CER(%) | $F0_{rmse}$(Hz) | LSD(dB) |
| CM-TTS | **3.11** | **1.29** | 37.58 | 4.58 | 19.88 | 6.88 | 29.64 | 4.64 | 16.04 | 10.50 | 25.75 | 4.53 |
| MM-TTS | 16.74 | 10.28 | 37.73 | 4.22 | 27.76 | 11.74 | 26.41 | **4.42** | 21.24 | 14.01 | 24.72 | **4.27** |
| MM-TTS$_{vae}$ | 4.87 | 2.34 | **36.57** | **4.15** | **15.13** | **4.36** | **26.28** | 4.59 | **14.45** | **9.51** | **24.24** | 4.36 |

Table2: Objective evaluation

| Model | First language : English | | | First language : Korean | | | First language : Japanese | | |
|---|---|---|---|---|---|---|---|---|---|
| | English | Korean | Japanese | English | Korean | Japanese | English | Korean | Japanese |
| CM-TTS | $2.71 \pm 0.12$ | $1.96 \pm 0.11$ | $2.16 \pm 0.11$ | $1.70 \pm 0.08$ | $2.75 \pm 0.10$ | $1.75 \pm 0.09$ | $1.77 \pm 0.10$ | $1.84 \pm 0.10$ | $2.93 \pm 0.12$ |
| MM-TTS | $2.93 \pm 0.12$ | $1.47 \pm 0.08$ | $1.91 \pm 0.11$ | $1.52 \pm 0.08$ | $2.15 \pm 0.10$ | $1.89 \pm 0.09$ | $1.96 \pm 0.12$ | $2.31 \pm 0.13$ | $2.98 \pm 0.12$ |
| MM-TTS$_{vae}$ | $\mathbf{3.13 \pm 0.12}$ | $\mathbf{2.15 \pm 0.12}$ | $\mathbf{2.20 \pm 0.12}$ | $\mathbf{2.13 \pm 0.09}$ | $\mathbf{3.03 \pm 0.10}$ | $\mathbf{2.34 \pm 0.11}$ | $\mathbf{2.30 \pm 0.12}$ | $\mathbf{2.66 \pm 0.12}$ | $\mathbf{3.15 \pm 0.12}$ |
| Recorded | $4.65 \pm 0.08$ | - | - | - | $4.94 \pm 0.03$ | - | - | - | $4.73 \pm 0.06$ |

Table3: Subjective evaluation

[1] R. Terashima et al., "Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation", Interspeech, 2022
[2] I. P. Association, "Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.", Cambridge University Press, 1999
[3] E. Song et al., "TTS-by-TTS 2: Data-selective augmentation for neural speech synthesis using ranking support vector machine with variational autoencoder", Interspeech, 2022
[4] H. Yoon et al., "Language model-based emotion prediction methods for emotional speech synthesis systems", Interspeech 2022

IEEE Advancing Technology for Humanity