2-1-22-TTS

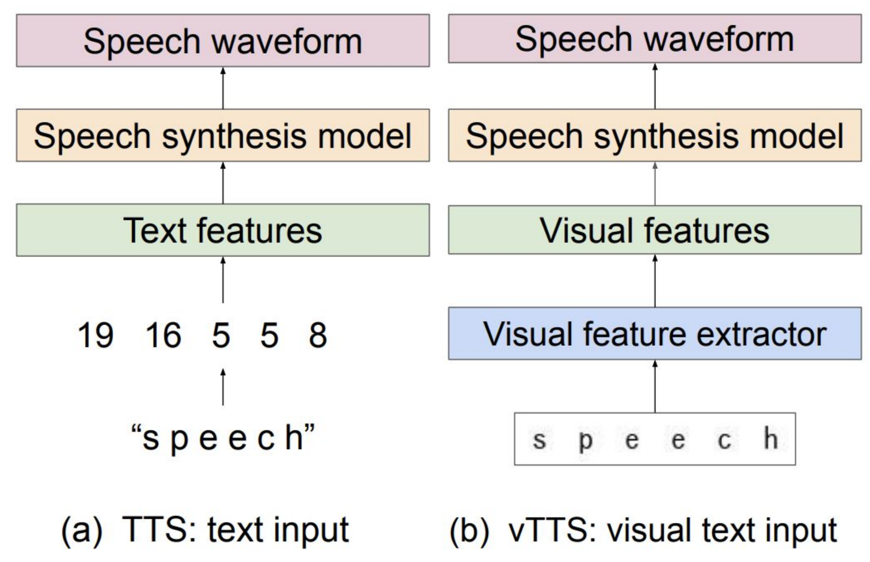
vTTS: visual-text to speech

Yoshifumi Nakano[1], Takaaki Saeki[1], **Shinnosuke Takamichi**[1], Katsuhito Sudoh[2], Hiroshi Saruwatari[1]
(1: The University of Tokyo, Japan. 2: Nara Institute of Science and Technology, Japan.)

Summary: synthesizing speech not from text (discrete symbols) but from visual text (text as an image)

- **Text is not a sequence of discrete symbols.**
  - Phonogram (e.g., Hangul)
    - A character representing a speech sound
    - Combination of sub-characters determines the reading
  - Emphasized word (e.g., underlined and **bold**) [1]
    - We read it emphatically.
  - Typefaces (e.g., in poster and comics) [2]
    - Utilizes to convey desired emotions to readers.
- **Text is an image! -> visual text (text as an image)**

- **Visual-text to speech (vTTS): a new task of speech synthesis**
  - Maps visual-text to speech.
  - We present an end-to-end mapping method.
- **Experiments**
  - Basic TTS (text to speech) vs. our vTTS
  - Transferring attributes in visual-text to speech
  - Robustness to OOV (out of vocabulary) characters



(a) TTS: text input (b) vTTS: visual text input

Methodology: end-to-end mapping from visual text to speech features

What visual texts do

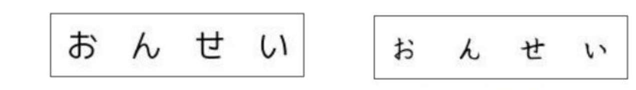
- Compositionality

강 (kang) = ㄱ (k) + ㅏ (a) + ㅇ (ng)

- Emphasis attribute

Underline Bold Italic

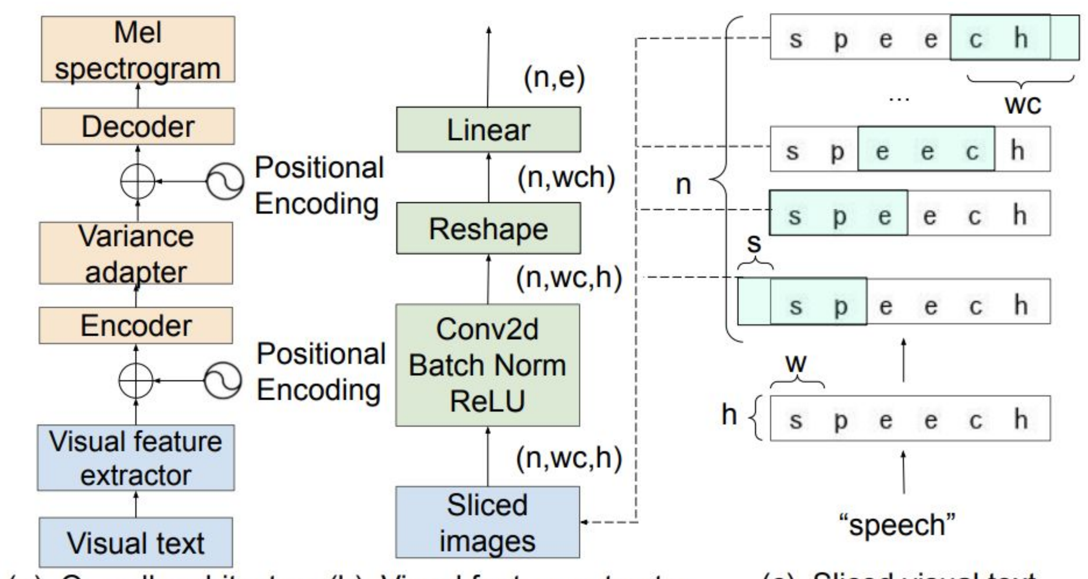
- Emotion attribute



Aiharahudemozikaisyo (sad) Koruri (joy)

- Visual-text conveys linguistic and para-linguistic information.
- Smallest units in speech synthesis
  - **Pixel (ours)** < byte [3] < phoneme < character < subword

vTTS model architecture

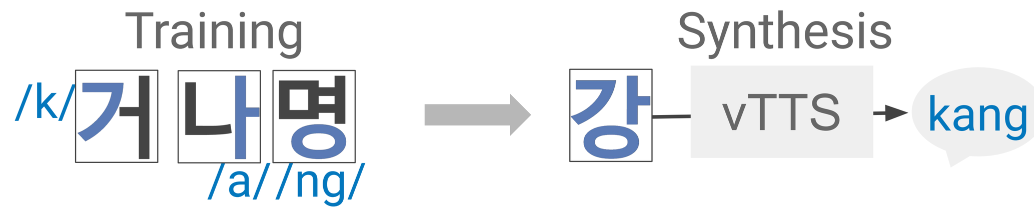


(a) Overall architecture (b) Visual feature extractor (c) Sliced visual text

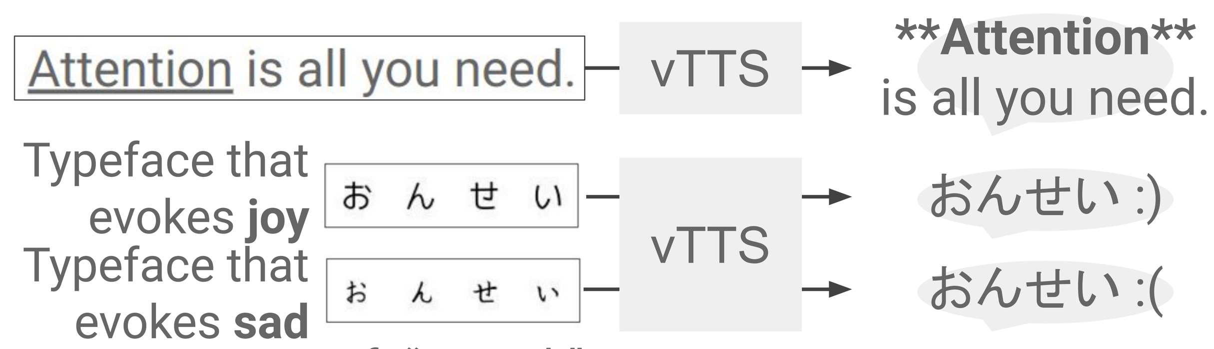
- **Visual text**
  - Artificially generated from text
    - Not realistic but good for benchmark
    - Monospace font
- **Visual feature extractor**
  - Extract visual features from visual text
- **FastSpeech 2 [4] encoder/decoder**
  - Non-autoregressive model

What the visual-feature extractor does

- **Compositionality of sub-characters**
  - In phonetic languages (e.g., Korean), combination of sub-characters determines the overall reading.
  - Even if OOV characters emerge, vTTS can predict the readings using the visual features.



- **Emphasis and emotion attributes**
  - The extractor will extract emphasis and typefaces.



Experimental evaluation

Experimental setup

Language	• Japanese (Hiragana) • Korean (Hangul) • English (Roman Alphabet)
Dataset	• 8.3 hours from JSUT (Japanese) [5] • + word-emphasized speech from JECS • + happy and sad speech from manga2voice [6] • 9.0 hours from KSS (Korean) [7] • 19 hours from LJSpeech (English) [8]
Model	• Character-input FastSpeech2 [5] (TTS) • Visual text-input model (vTTS) (All the models are mono-lingual.)

Transferring emphasis

- **"Which word is emphasized?"**
  - Listener listens to synthetic speech and answer the emphasized word.
  - Emphasis is accurately transferred.

Speech (Ja)		Accuracy
Ground truth		**0.933**
Underline	Attention is all ...	**0.933**
Bold	**Attention** is all ...	0.898
Italic	*Attention* is all ...	0.877
No effect	Attention is all ...	0.381 ~ 0.505

Transferring emotion

- **"Which emotion is perceived?"**
  - Listener listens to synthetic speech and answer the perceived emotion.
  - Emotion is accurately transferred.

Confusion matrix (Ja)	Happy (perceived)	Sad (perceived)
Happy (true) おんせい	**0.795**	0.205
Sad (true) おんせい	0.114	**0.886**

TTS vs. vTTS: comparison of naturalness

- **5-point mean opinion score (MOS) on naturalness**
  - Language-wise evaluation

Lang.	TTS	vTTS (window) c=1 s p e e c h	c=3 s p e e c h	c=5 s p e e c h
Ja	3.45 ± 0.09	3.41 ± 0.09	3.46 ± 0.09	3.49 ± 0.10
Ko	**3.04 ± 0.16**	**3.55 ± 0.15**	3.18 ± 0.15	3.01 ± 0.15
En	3.72 ± 0.10	3.69 ± 0.10	3.70 ± 0.11	3.71 ± 0.10

- **TTS vs. vTTS**
  - Comparable in Ja and En (no significant difference)
  - vTTS is better in Ko (significant difference)
- **Effect of window size *c***
  - Naturalness improves as c increases in Ja and En.
  - *c* = 1 is the best in Ko (due to the number of phonemes expressed by one character?)

Robustness to OOV character

- **Three test sets**
  - "**in-vocab**" consists of characters appearing more than 3 times in training data.
  - "**rare**" includes appearing less than 3 times in the training data.
  - "**OOV**" includes OOV characters.
- **Evaluation (Korean speech only)**
  - 5-point MOS on naturalness by native speakers
  - Character error rate (CER) of transcription by native speakers
  - vTTS is more robust to OOV (= degradation by OOV is small) than TTS.

MOS (Δ: decrease from "in-vocab.")

	in-vocab	rare (Δ)	OOV (Δ)
TTS	3.29 ± 0.16	2.32 ± 0.16 (**−0.97**)	2.31 ± 0.20 (**−0.98**)
vTTS	3.58 ± 0.13	3.12 ± 0.16 (**−0.46**)	2.95 ± 0.21 (**−0.63**)

CER (Δ: decrease from "in-vocab.")

	in-vocab	rare (Δ)	OOV (Δ)
TTS	0.120	0.194 (**+0.074**)	0.255 (**+0.135**)
vTTS	0.080	0.114 (**+0.034**)	0.163 (**+0.083**)

Future direction

- vTTS from real image, e.g., posters, comics (manga), and other in-the-wild images.

Reference
[1] Strobelt et al., IEEE TVCG, 2016.
[2] S. Choi et al., AltMM, 2016.
[3] B. Li et al., ICASSP, 2019.
[4] Y. Ren et al., ICLR, 2021.
[5] R. Sonobe et al., AST, 2019.
[6] S. Takamichi et al., ASJ, 2020.
[7] https://kaggle.com/bryanpark/ korean-single-speaker-speech-dataset
[8] https://keithito.com/LJ-Speech-Dataset/