SONY

Polyphone disambiguation and accent prediction using pre-trained language models in Japanese TTS front-end

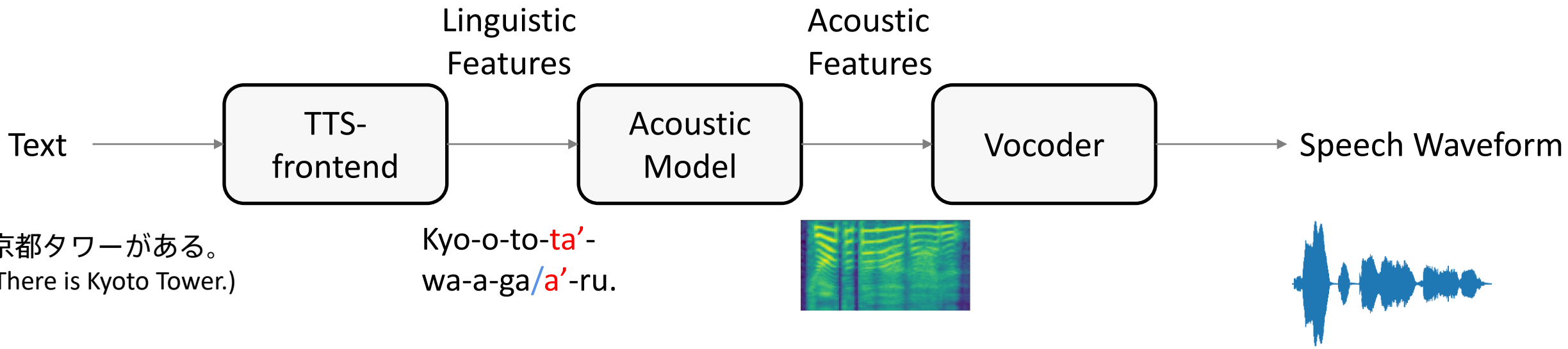
IEEE ICASSP 2022

Rem Hida, Masaki Hamada, Chie Kamada
Emiru Tsunoo, Toshiyuki Sekiya, Toshiyuki Kumakura
Remu.Hida@sony.com

Sony Group Corporation
R&D Center

TTS frontend

TTS frontend converts text to phonetic symbol sequences.



Highly language dependent
Japanese has mainly two characteristics.

Japanese Characteristics related to TTS

- Japanese has a variety of character types and its pronunciation.
  - Some Kanji have multiple candidate pronunciations corresponding to different meanings.

このカレーは**辛い**。(This curry is **spicy**.)
- ka-ra-i (spicy)
- tsu-ra-i (hard)

➡ **Polyphone disambiguation**

- Japanese is pitch (High/Low)-accent language.
  - Some words have the same pronunciation but different accents and meanings.


chopsticks — ha / shi


bridge — ha / shi

  - The pitch accent is represented by the **accent phrase boundary** and the **accent nucleus position**.

Word		京都	タワー	が	ある
Accent	High / Low	kyo / o – to – ta / wa – a – ga			a / ru

➡ **Accent prediction**

Wrong pronunciation & accent lead wrong comprehension.
Japanese TTS system requires "**polyphone disambiguation (PD)**" and "**accent prediction (AP)**."

Motivation

Pronunciation and accent depend on context.

PD: このカレーはとても**辛い (ka-ra-i)**。 体調が悪くてとても**辛い (tsu-ra-i)**。

This curry is very **spicy**. I'm sick and it's very **hard**.

AP:

京都 (Kyoto)	タワー (tower)	上空 (above)
kyo'-o-to	ta'-wa-a	jo-o-ku-u
kyo-o-to-ta'-wa-a		jo-o-ku-u

Semantic relationship

However, existing methods only utilize local context.

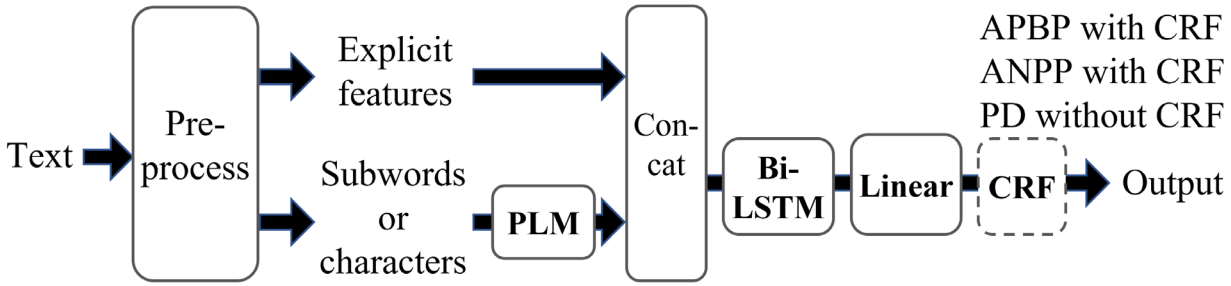
PD: KyTea[Neubig+,10] (pointwise prediction)

AP: TASET[Suzuki+,17] (linear-chain CRF)

How to take "longer/rich context" into account?
-> Using **Pre-trained Language Models**.

Japanese TTS-frontend with Pretrained Language Models (PLMs)

Model:



Features

Explicit(EF):

features derived from morphological analysis

Implicit(PLM):

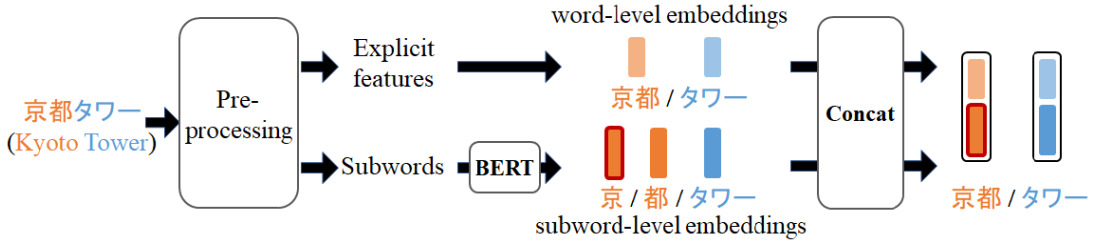
features from Pretrained Language Models

BERT: subword based masked language model

Flair: character based bidirectional encoder

Word	京都	タワー	が	ある	
POS	Noun	Noun	Particle	Verb	Explicit Features (EFs)
Original pronunciation	kyo-o-to	ta-wa-a	ga	a-ru	
Accent nucleus position of each word	kyo \ o – to	ta \ wa – a	ga	a \ ru	
Other features	・・・	・・・	・・・	・・・	

Explicit and implicit features are concatenated and input into BiLSTM.



Dataset for Experiments

- Polyphone disambiguation

  Focus on 92 frequently used polyphonic words

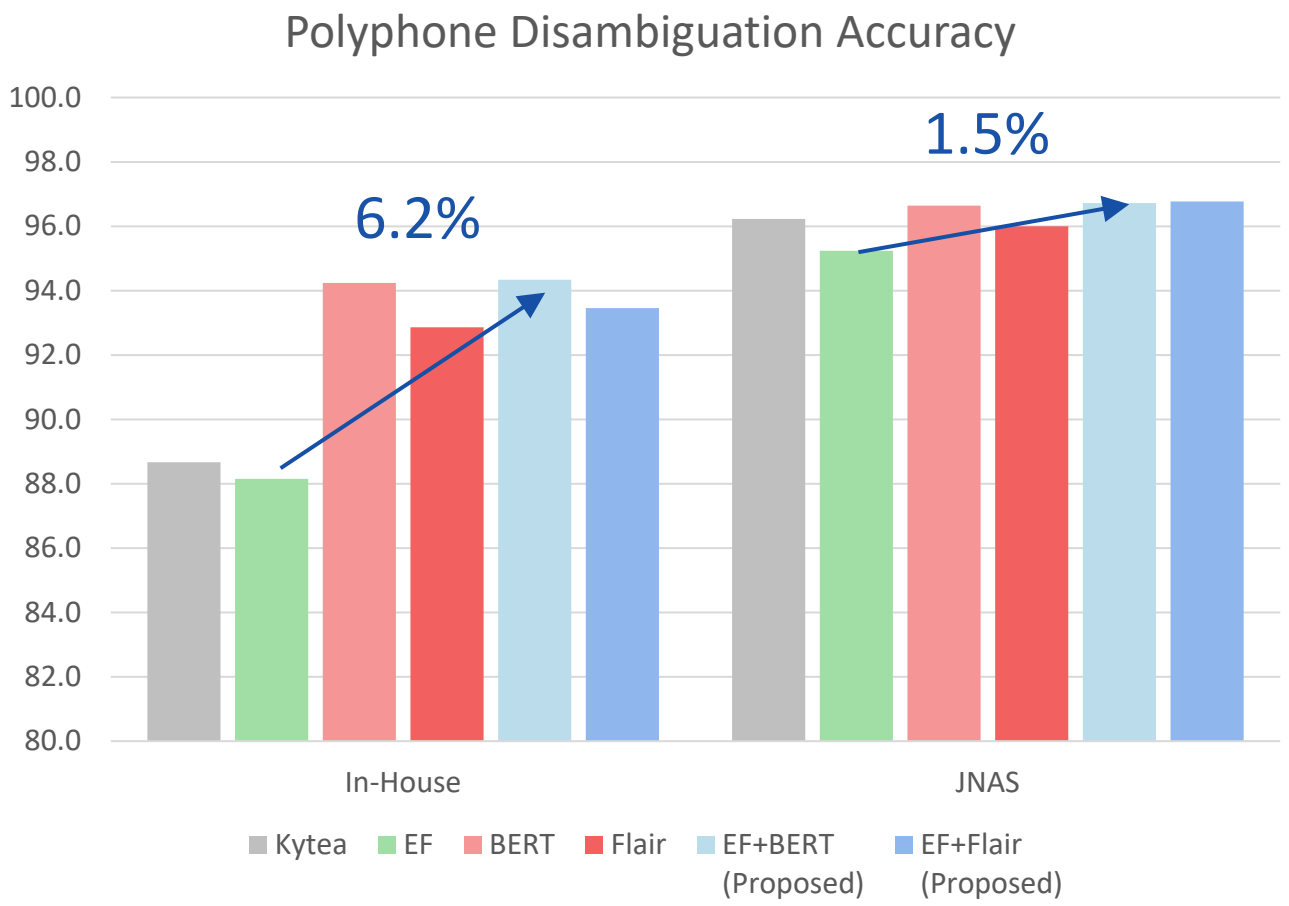
	#sentence	usage	Source
In-house	39,353 (24,117 / 5,156 / 10,080)	Train/dev/test	Wikipedia/TV captions/ novels/CSJ/JSUT
Public (JNAS)	5,642	test	JNAS

- Accent Prediction

	#sentence	usage	Source
In-house	9,497 (7,768 / 864 / 865)	Train/dev/test	TV caption
Public (JSUT)	5,000	test	JSUT

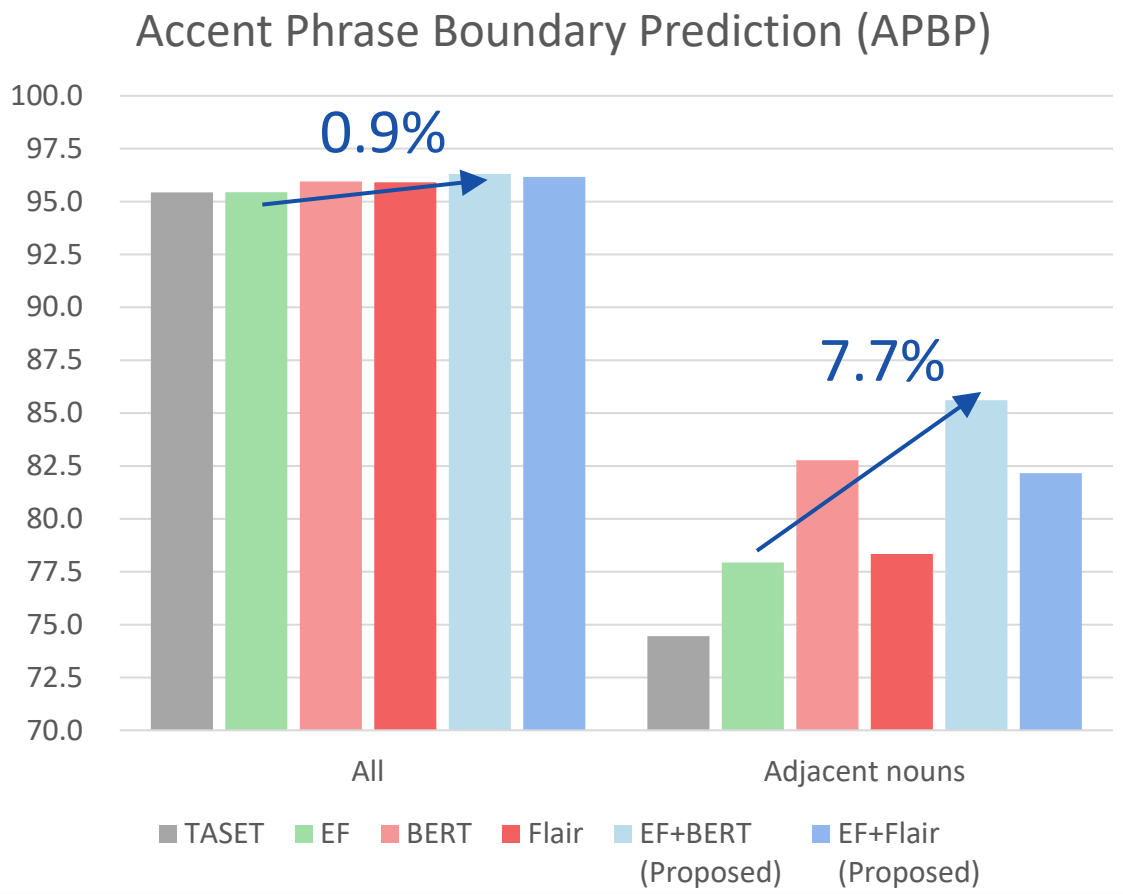
1: K. Maekawa, “Corpus of spontaneous Japanese: its design and evaluation,” in ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition, 2003, pp. 7–12.
2 : R. Sonobe, S. Takamichi, and H. Saruwatari, “JSUT corpus: free largescale Japanese speech corpus for end-to-end speech synthesis,” arXiv preprint arXiv:1711.00354, 2017.
3: ASJ Japanese Newspaper Article Sentences Read Speech Corpus (JNAS), http://research.nii.ac.jp/src/JNAS.html.

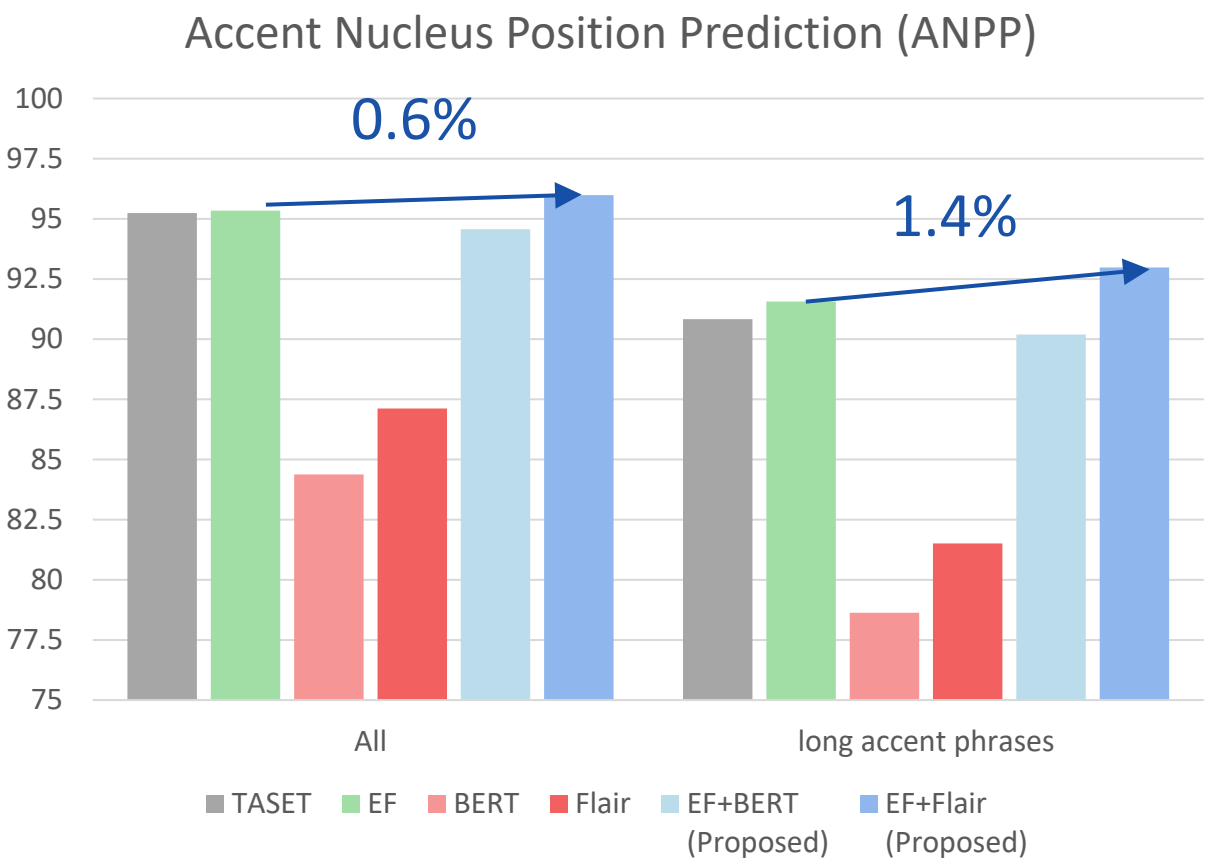
Comparison w/ & w/o PLMs on Polyphone disambiguation



Bi-LSTM w/ EF+BERT improves by 6.2/1.5% from only EF.

Comparison w/ & w/o PLMs on Accent Prediction





Bi-LSTM w/ EF+PLM improves by 0.9/7.7% on APBP, 0.6/1.4% on ANPP from only EF.
BERT for APBP, Flair for ANPP

TTS quality subjective evaluation

TTS settings

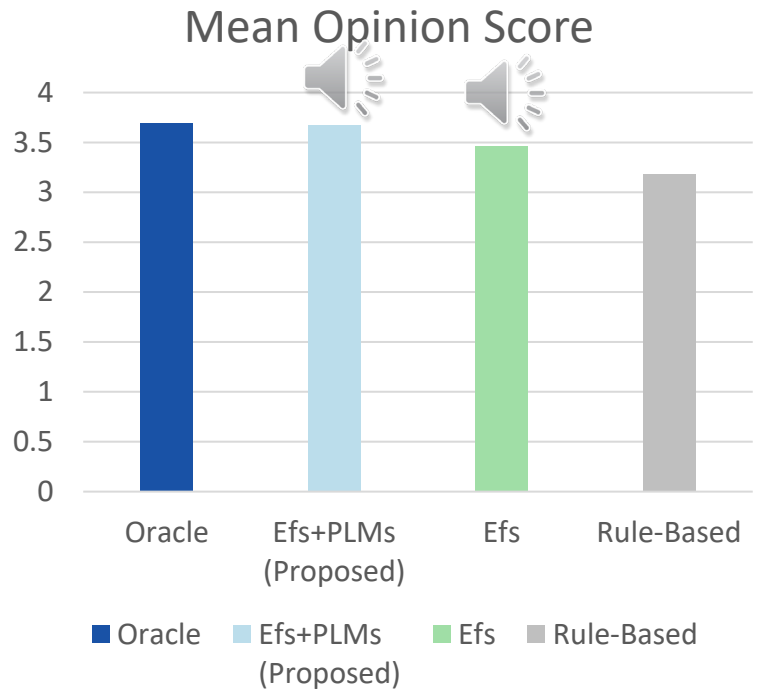
- Acoustic model: Tacotron2 w/ Global Style
- Token[Shen+,18. Wang+,18]
- Vocoder: Parallel WaveGAN [Yamamoto+, 20]
- Training data: Sub-corpus of JSUT for both models

Evaluator

- 30 native Japanese Speakers

Evaluation text

- 25 utterance samples from in-house data



Proposed method achieved almost the same speech quality as Oracle.

Summary

We proposed the method which incorporates implicit/explicit features in PD/AP.

- The combination of explicit and implicit features improves both PD/AP performance.
- Methods showed better performance on MOS than conventional TTS-frontend.
- The effectiveness of PLMs type (BERT/Flair) depends on tasks.

Future Work

・Using pre-trained model from both of grapheme & phoneme

References

J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *NAACL-HLT2019*, 2019.

G. Neubig and S. Mori, “Word-based partial annotation for efficient corpus construction,” in LREC, 2010, pp. 2723–2727.

M. Suzuki et al., “Accent sandhi estimation of Tokyo dialect of Japanese using conditional random fields,” IEICE Trans. Inf. & Syst., vol. 100, no. 4, pp. 655–661, 2017.

J. Shen et al., “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” ICASSP, 2018,

Y. Wang et al., “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” ICML, 2018,

R. Yamamoto et al., “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” ICASSP, 2020