SLP-P15.4

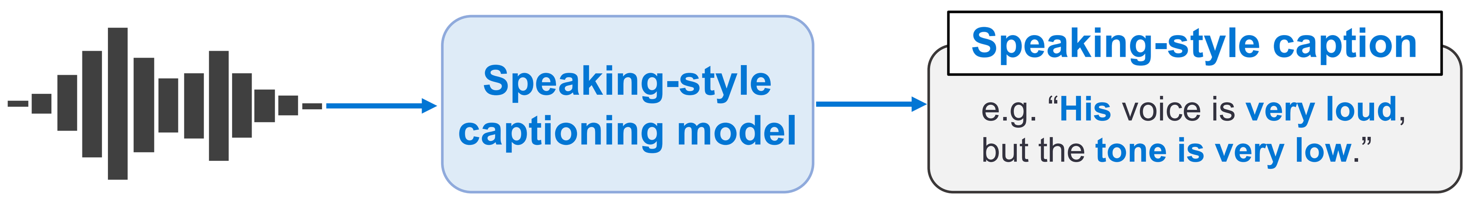
StyleCap: Automatic Speaking-Style Captioning from Speech Based on Speech and Language Self-supervised Learning Models

Kazuki Yamauchi*, Yusuke Ijima†, Yuki Saito*
*The University of Tokyo, Japan, †NTT Corporation, Japan

Demo page

Contributions: speaking-style captioning and StyleCap

- **Contribution 1: new task, speaking-style captioning**
  - Describe ***speaking-style*** in speech **in natural language**
  - **Go beyond the limitations by pre-defined discrete labels**



- **Related works: para-/non-linguistic information recognition**
  - Classify speech into **pre-defined categories**
  - Desire **Human-interpretable reasoning**

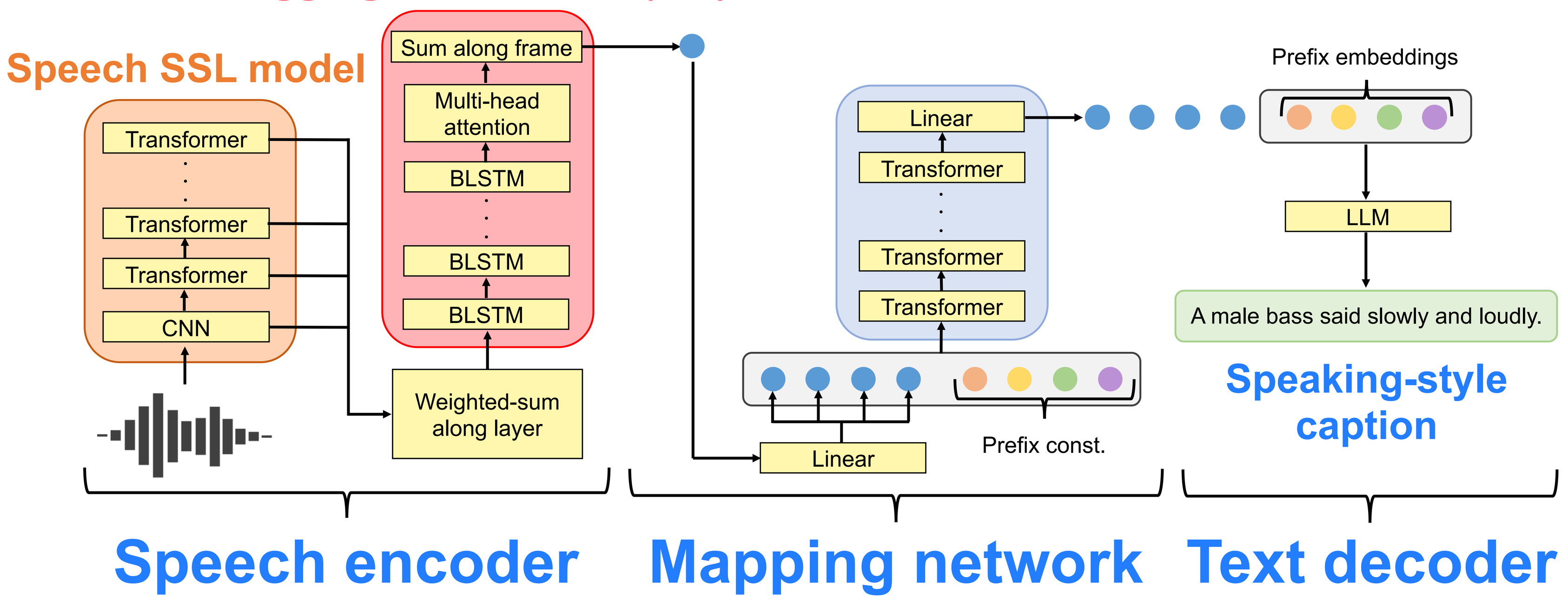
- **Contribution 2: speaking-style captioning model, StyleCap**
  - Leverage speech and language **SSL models**
  - Introduce LLM-based data augmentation by **sentence rephrasing**

- **Related works: image/audio captioning**
  - Describe ***content*** **information** in natural language
  - e.g. **ClipCap** [1], leveraging SSL for image captioning

Method: StyleCap, leveraging SSL models and LLM-based data augmentation

DNN architecture of StyleCap

- **SSL-based speech encoder**
  - Compress SSL features into a fixed-length vector using AM
- **Q-former-based mapping network**
  - Map the fixed-length vector to the word embedding space
- **LLM-based text decoder**
  - Generate a caption from the mapping network output by LLM



Data augmentation by sentence rephrasing

- **Challenge: one-to-many problem of style prediction**
  - Not uniquely determined descriptions to express speaking-style

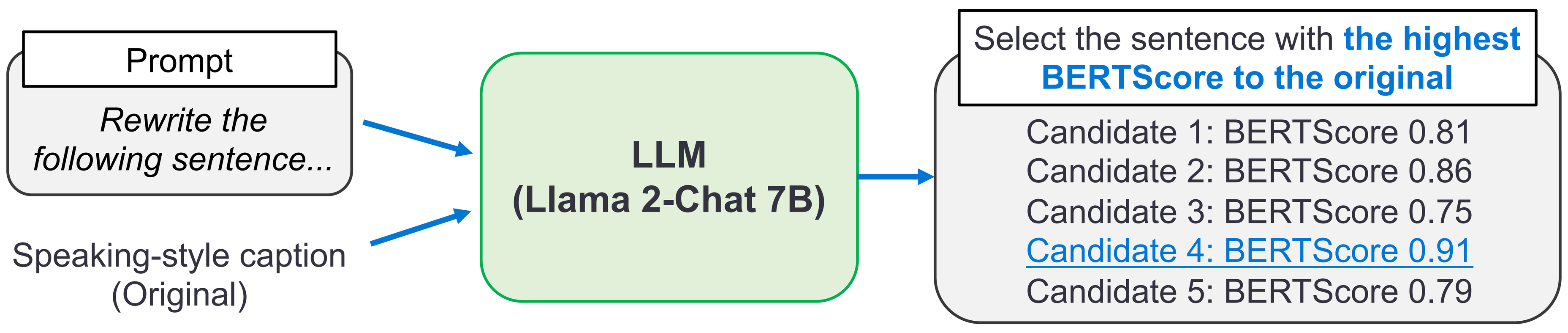
  → **Data augmentation by sentence rephrasing using LLM** to increase the diversity of description
- **Problem: change in meaning due to rephrasing error by LLM**

  → **Filtering by BERTScore** [2] **to select appropriate candidates**
- **Example of rephrasing:**

His sound height is normal, but the speed is very fast, and the **volume is very low.**

↓ **Rephrased by LLM (Llama 2-Chat 7B model** [3]**)**

Despite his normal height, his sound is incredibly fast and **surprisingly quiet.**



Experimental evaluation

Experimental conditions

- **Dataset: PromptSpeech** [4]
  - Various **(speech, style prompt) pairs** of data
  - Speech in **LibriTTS corpus** [5] annotated with style prompt
  - **Style factor: gender, pitch, speed, volume**
  - Various speakers/utterances: 1,191/26,588
- **Model configuration of StyleCap**
  - Speech encoder: **WavLM BASE+** [6] / **mel-spec.** / **x-vector** [7]
  - Mapping network: Transformer encoder x 8
  - Text decoder: **GPT-2 (125M params)** [8] / **Llama 2 (7B params)**

Experimental results

- **Metrics: METEOR** [9] **(M), BERTScore (BS), Distinct-1** [10] **(D1)**

Speech encoder	Text decoder	w/o Sentence Rephrasing M(↑)	BS(↑)	D1(↑)	w/ Sentence Rephrasing M(↑)	BS(↑)	D1(↑)
Mel-spec. + AM	GPT-2	0.357	0.827	0.020	0.334	0.822	0.021
x-vec.	GPT-2	0.255	0.800	0.013	0.237	0.798	0.013
WavLM + AM	GPT-2	**0.410**	**0.839**	**0.022**	0.439	0.848	0.022
Mel-spec. + AM	Llama 2	0.332	0.821	0.022	0.327	0.818	0.024
x-vec.	Llama 2	0.239	0.799	0.016	0.237	0.769	0.014
WavLM + AM	**Llama 2**	**0.469**	**0.855**	**0.023**	**0.479**	**0.857**	**0.027**

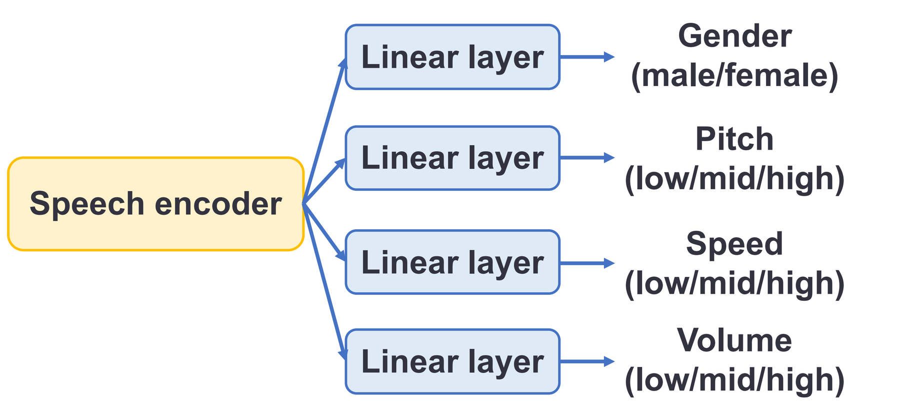
1. **Speech encoder validity: WavLM + AM > others**
2. **Text decoder validity: Llama 2 > GPT-2**
3. **Sentence Rephrasing: Improved diversity in particular**

Analysis of StyleCap behavior

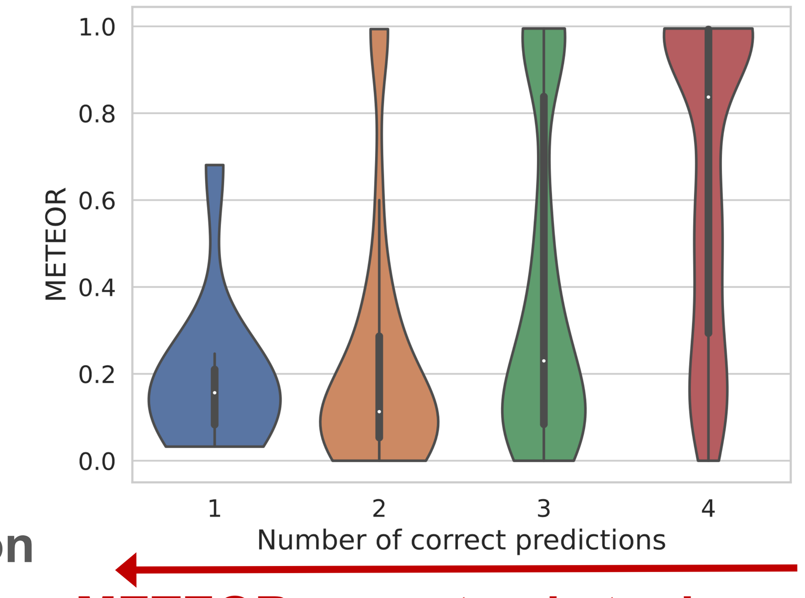
- **Style factor classification**
  - Predict style factors from each trained speech encoder output
    - 2 classes for gender (male/female)
    - 3 classes for pitch, speed, volume (low/mid/high)

Speech encoder	gender	pitch	speed	volume	Ave.
Mel-spec. + AM	93.8	62.1	67.8	54.7	**69.6**
x-vec.	94.0	40.5	44.0	49.4	**57.0**
WavLM + AM	91.0	61.0	85.2	69.9	**76.8**

Classification performance: WavLM + AM > others



Overview of style factor classification



METEOR score tends to degrade

- **METEOR distribution by the number of correct**
  - Captioning performance (i.e., METEOR score) **degrades** as the number of correct predictions **decreases**

→ **Capturing adequate para-/non-linguistic information** by **the speech encoder is an essential factor**

Future direction

- **Adapt to other para-/non-linguistic information (e.g. emotion)**
- **Dataset construction for more diverse speaking-style caption**

Acknowledgments: This work was supported by JST, ACT-X Grant Number JPMJAX23CB, Japan.

Reference

[1] R. Mokady et al., arXiv preprint arXiv: 2111.09734, 2021. [2] T. Zhang et al., in Proc. ICLR, 2020. [3] H. Touvron et al., arXiv preprint arXiv:2307.09288, 2023. [4] Z. Guo et al., in Proc. ICASSP, 2023.
[5] H. Zen et al., in Proc. INTERSPEECH, 2019. [6] S. Chen et al., IEEE JSTSP, 2022. [7] D. Snyder et al., in Proc. ICASSP, 2018. [8] A. Radford et al., 2019. [9] S. Banerjee et al., in Proc. ACL, 2005. [10] J. Li et al., in Proc. ACL, 2016.

