

2024-03-08 日本音響学会第151回 (2024年春季) 研究発表会 3-2-14

StyleCap:

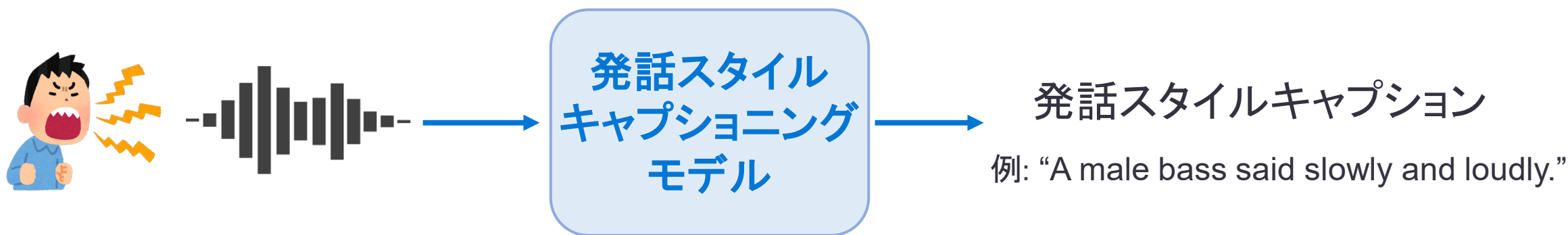
音声および言語の自己教師あり学習モデルに基づく
音声の発話スタイルに関するキャプション生成

☆ 山内 一輝 (NTT/東大院・情報理工), 井島 勇祐 (NTT), 齋藤 佑樹 (東大院・情報理工)

はじめに: 本研究の貢献

○貢献1: 「**発話スタイルキャプション**」という新たなタスクを提案

- 音声に含まれる発話スタイル・感情表現を**自然言語で記述**
- 事前に定義された離散ラベル (例: 感情) の予測に縛られないスタイル表現を生成



○貢献2: 発話スタイルキャプション生成のためのモデル「**StyleCap**」を提案

- 画像キャプション手法 (ClipCap [Mokady et al., 2021]) を本タスク向けに再構築
- より良い特徴抽出のために**音声と言語の自己教師あり学習 (SSL) モデルを活用**
- 大規模言語モデル (LLM) に基づく **Sentence Rephrasing** によるデータ拡張も提案

関連研究

○パラ言語/非言語情報認識

- 入力された音声から感情や話者性などのパラ言語/非言語情報を推定
- 高い認識精度だけでなく, 人間が**解釈可能な理由付け**も得られることが望ましい
 - 従来手法: **事前に定義されたカテゴリへの分類や強度の推定**が主目的

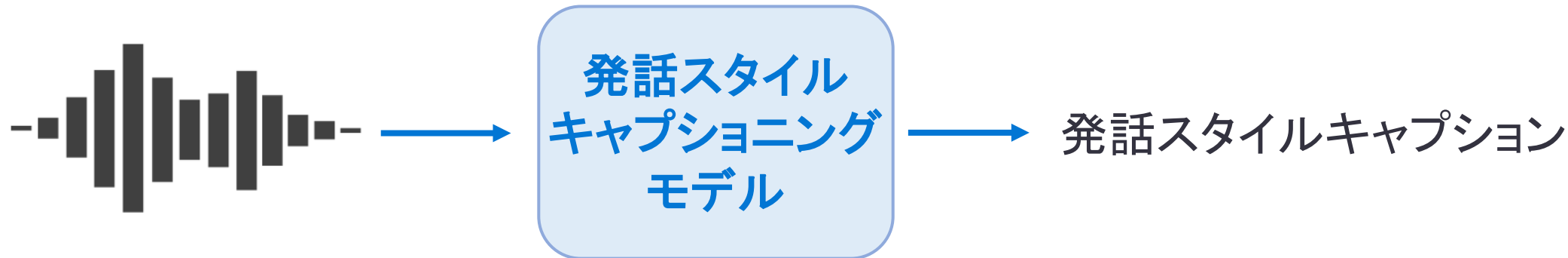
→ **パラ言語/非言語情報の自然言語による記述が解決策になり得る**

○画像/オーディオキャプション

- 画像/オーディオデータの内容情報を自然言語で記述
- 例: 画像キャプションのための **ClipCap** [Mokady et al., 2021]
 - 画像と言語の SSL モデルを活用し, 高精度なキャプションを実現

→ **発話スタイルキャプションにも音声と言語の SSL モデルを活用**

発話スタイルキャプショニングの概要



例:

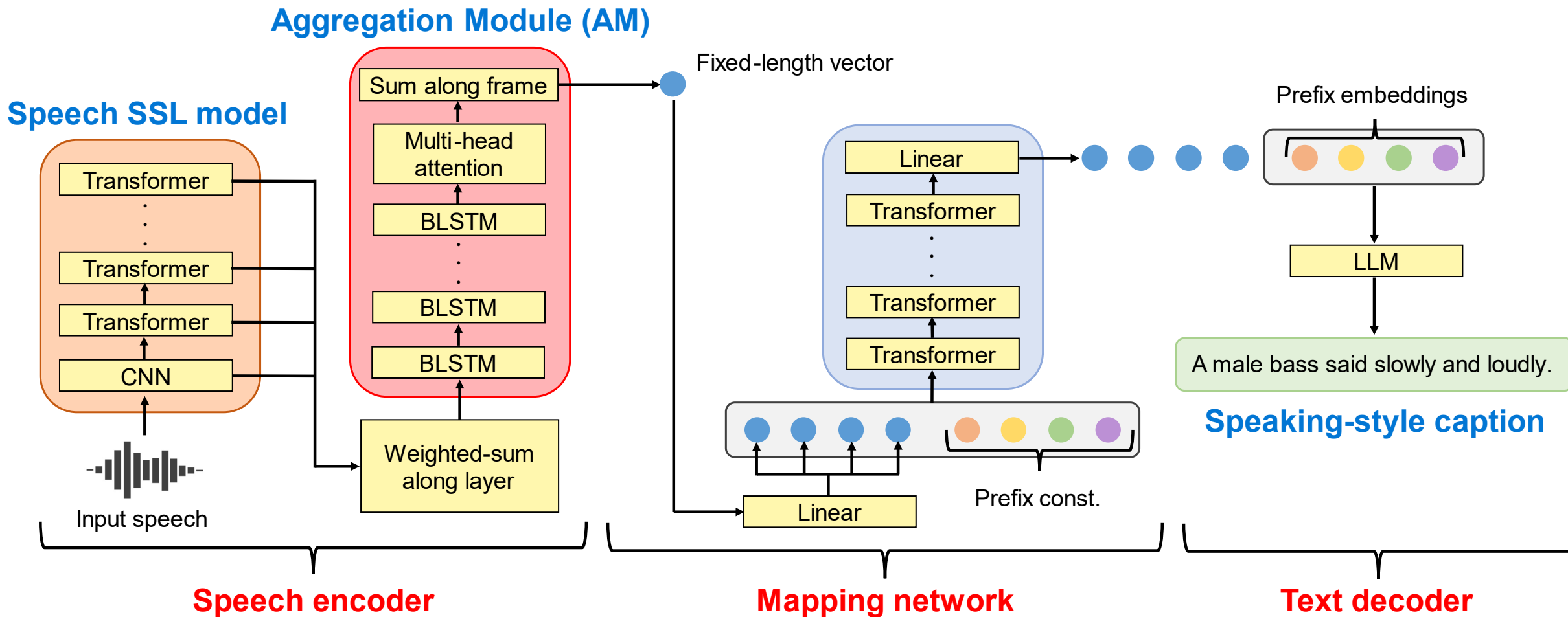


“His voice is very loud,
but the tone is very low.”

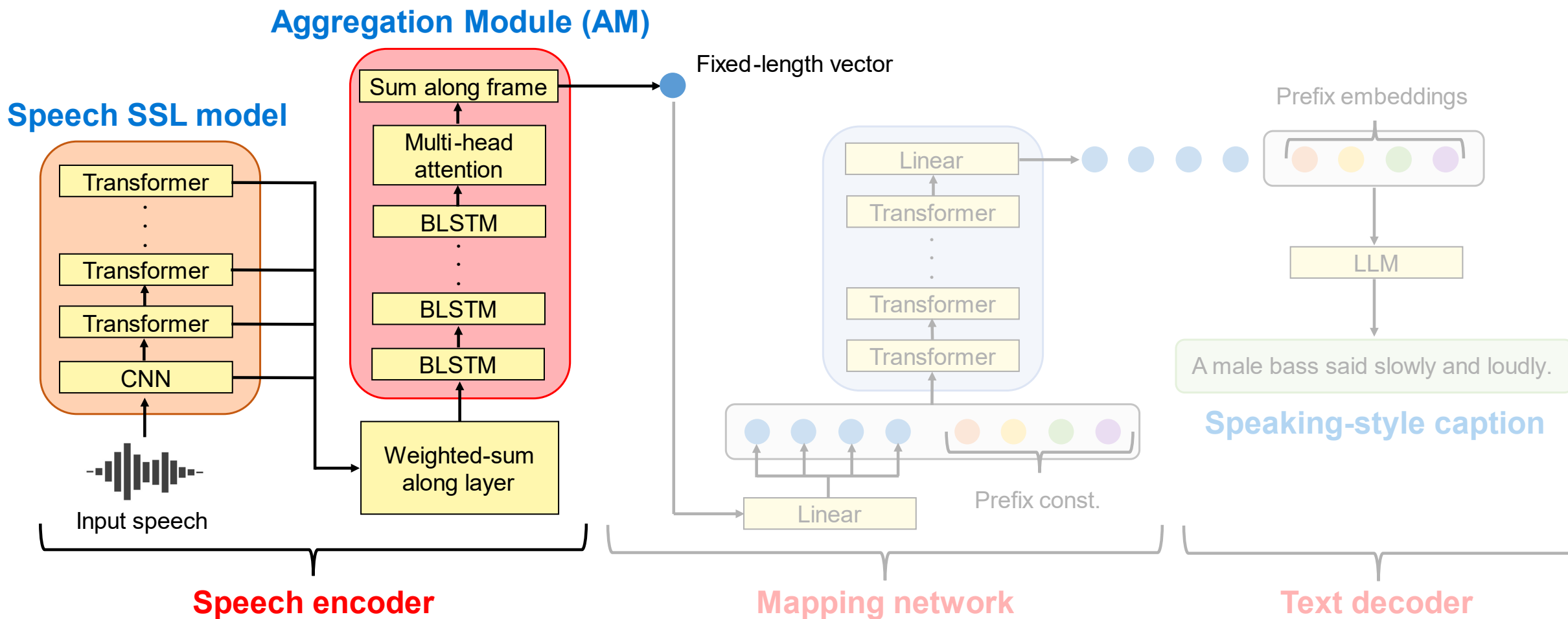
○本研究で用いたデータセット: PromptSpeech コーパス [Guo et al., 2023]

- 様々な (音声, 発話スタイルに関する記述文) のペアデータにより構成
- 自然言語で発話スタイルを制御可能なテキスト音声合成モデルに向けて構築
 - 主な制御対象: 話者の性別, 音高, 話速, 音量
 - LibriTTS コーパス [Zen et al., 2019] に含まれる音声に発話スタイルに関する記述文を付与

StyleCap のアーキテクチャ

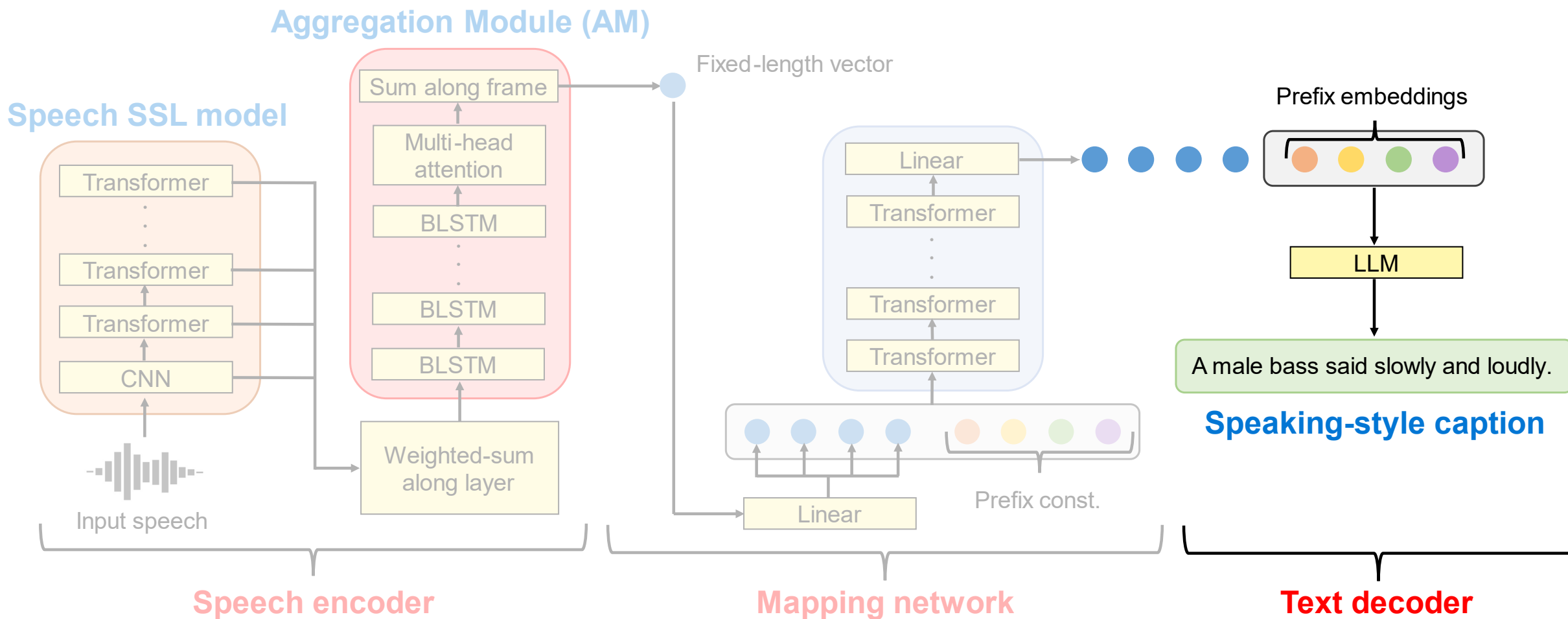


StyleCap のアーキテクチャ



音声から SSL 特徴量を抽出し, Aggregation Module (AM) で固定長ベクトルに圧縮

StyleCap のアーキテクチャ



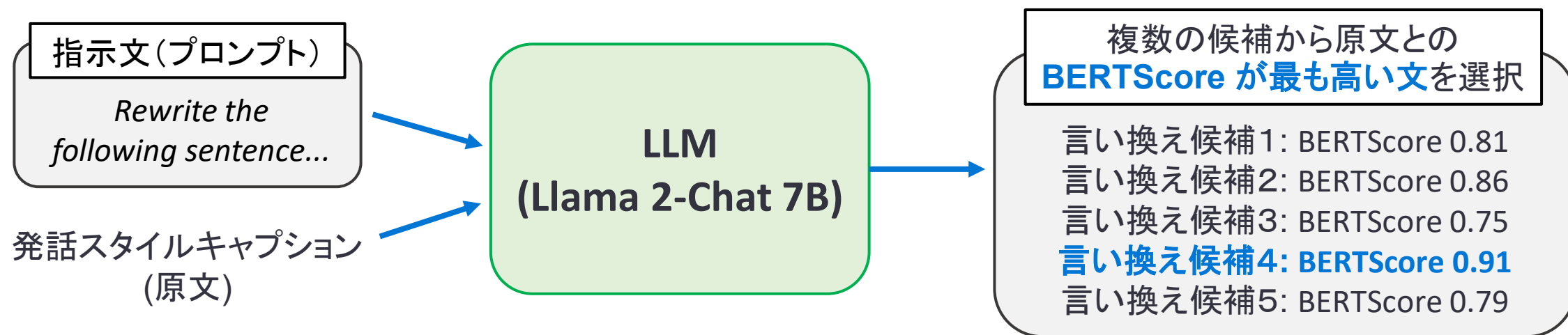
Mapping network の出力を LLM に入力し, 音声発話スタイルのキャプションを生成

Sentence Rephrasingによるデータ拡張

○発話スタイルキャプションの難しさ: **スタイル予測の one-to-many 問題**

□1つの音声に対する発話スタイル表現は一意に定まらず, 複数の解が存在

→ **LLMによる Sentence Rephrasing でスタイル記述の多様性を増加**



例: 原文: “His sound height is normal, but the speed is very fast, and the **volume is very low.**”
変換後: “Despite his normal height, his sound is incredibly fast and **surprisingly quiet.**”

實驗的評價

実験条件 (詳細は原稿参照)

○StyleCap のモデル設定

- Speech encoder: (**WavLM BASE+** + AM) or (メルスペクトログラム + AM) or x-vector
 - Aggregation module (AM): 固定長ベクトル化のためのモジュール
- Mapping network: Transformer encoder x 8
- Text decoder: **GPT-2** (125M params) or **Llama 2** (7B params)

○データセット: PromptSpeech コーパス [Guo et al., 2023]

- 話者数/発話数: 1,191名/約26,000発話
- 学習/検証/評価データ数: 24,953/857/778発話

○DNN 学習の条件

- バッチサイズ/エポック数: 16/20(データ拡張無しなら10)
- 損失関数: Ground-truth キャプションと生成キャプション間の Cross-entropy loss

提案手法の構成要素に関する客観評価

○評価指標: METEOR (M), BERTScore (BS), Distinct-1 (D1) (他は原稿参照)

		w/o Sentence Rephrasing			w/ Sentence Rephrasing		
Speech encoder	Text decoder	M(↑)	BS(↑)	D1(↑)	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

提案手法の構成要素に関する客観評価

○評価指標: METEOR (M), BERTScore (BS), Distinct-1 (D1) (他は原稿参照)

		w/o Sentence Rephrasing			w/ Sentence Rephrasing		
Speech encoder	Text decoder	M(↑)	BS(↑)	D1(↑)	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

Speech encoder の有効性: **WavLM + AM** > others

提案手法の構成要素に関する客観評価

○評価指標: METEOR (M), BERTScore (BS), Distinct-1 (D1) (他は原稿参照)

		w/o Sentence Rephrasing			w/ Sentence Rephrasing		
Speech encoder	Text decoder	M(↑)	BS(↑)	D1(↑)	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

Speech encoder の有効性: **WavLM + AM** > **others**

提案手法の構成要素に関する客観評価

○評価指標: METEOR (M), BERTScore (BS), Distinct-1 (D1) (他は原稿参照)

		w/o Sentence Rephrasing			w/ Sentence Rephrasing		
Speech encoder	Text decoder	M(↑)	BS(↑)	D1(↑)	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

Text encoder の有効性: **Llama 2** > **GPT-2**

提案手法の構成要素に関する客観評価

○評価指標: METEOR (M), BERTScore (BS), Distinct-1 (D1) (他は原稿参照)

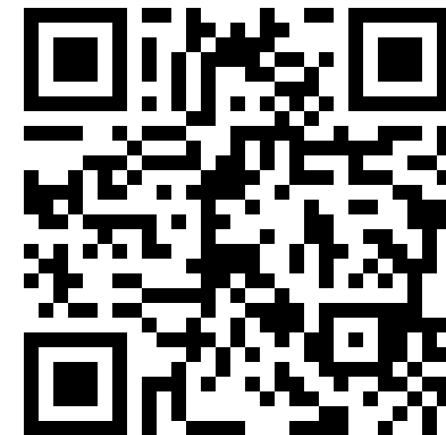
		w/o Sentence Rephrasing			w/ Sentence Rephrasing		
Speech encoder	Text decoder	M(↑)	BS(↑)	D1(↑)	M(↑)	BS(↑)	D1(↑)
Mel-spec. + AM	GPT-2	0.357	0.827	0.020	0.334	0.822	0.021
x-vec.	GPT-2	0.255	0.800	0.013	0.237	0.798	0.013
WavLM + AM	GPT-2	0.410	0.839	0.022	0.439	0.848	0.022
Mel-spec. + AM	Llama 2	0.332	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	<u>0.023</u>	0.479	0.857	<u>0.027</u>

Sentence Rephrasing: 生成キャプションの多様性を特に改善

サンプル

Speech: 

デモページはこちら 



Ground-truth: His tone is so high, the volume is very large, and he speak very fast.

Speech encoder	Text decoder	Generated caption
Mel-spec. + AM	GPT-2	His sound height is really high, the volume is very large, and he speaks quickly.
x-vec.	GPT-2	Please generate a man with a loud voice to <u>say slowly</u> .
WavLM + AM	GPT-2	His sound height is really high, the volume is very large, and he speaks quickly.
Mel-spec. + AM	Llama 2	His tone is so high, the volume is very large, and he speaks quickly.
x-vec.	Llama 2	His sound height is really high, the volume is very large, and he speaks quickly.
WavLM + AM	Llama 2	His tone is so high, the volume is very large, and he speak very fast.

x-vector は発話スタイルに関する情報はあまり保持していないため、本タスクには不適切

まとめ

○貢献1: 発話スタイルキャプションの提案

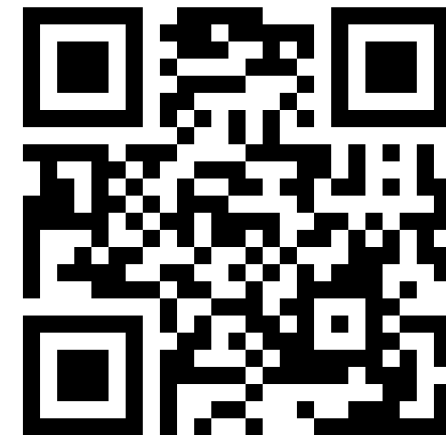
- 音声に含まれる発話スタイル・感情表現を自然言語で記述
- 事前に定義された離散ラベル (例: 感情) の予測に縛られない

○貢献2: StyleCap の提案 & 有効性の評価

- SSL モデル由来の音声表現ベクトルと、表現力の高い LLM を用いた StyleCap が最も高い精度の発話スタイルキャプションを実現
- Sentence Rephrasing によるデータ拡張を用いることで、生成される発話スタイルキャプションの精度と多様性が向上

○今後の課題

- 他のパラ言語/非言語情報 (感情, 心的状態など) のキャプションへの適応
- より多様な発話スタイルキャプションに向けたデータセット構築



ICASSP2024版の
論文はこちら

Appendix: ベースラインモデルとの比較

○ベースラインモデル: Transformer encoder-decoder

□入力: メルスペクトログラム

□Encoder: Transformer encoder x 12, Decoder: Transformer decoder x 6

○評価指標: METEOR (M), BERTScore (BS), Distinct-1 (D1) (他は原稿参照)

	w/o Sentence Rephrasing			w/ Sentence Rephrasing		
Model	M(↑)	BS(↑)	D1(↑)	M(↑)	BS(↑)	D1(↑)
Transformer encoder-decoder	0.320	0.817	0.019	0.303	0.814	0.018
StyleCap w/ WavLM, Llama 2	0.469	0.855	0.023	0.479	0.857	0.027

Model の有効性: **StyleCap** > **Transformer encoder-decoder**

Appendix: 実験結果詳細(データ拡張なし)

○評価指標:

□B@4: BLEU4, R: ROUGE, M: METEOR, BS: BERTScore, C: CIDEr-D, S: SPICE scores

Model	Speech encoder	B@4(↑)	R(↑)	M(↑)	BS(↑)	C(↑)	S(↑)	distict-1(↑)	distict-2(↑)
Transformer encoder-decoder	Mel-spectrogram	0.163	0.352	0.320	0.817	2.171	0.273	0.019	0.049
	x-vector	0.096	0.269	0.248	0.799	1.289	0.209	0.015	0.039
	WavLM	0.253	0.475	0.456	0.850	3.239	0.419	0.022	0.064
StyleCap w/ GPT-2	Mel-spectrogram + AM	0.178	0.381	0.357	0.827	2.295	0.316	0.020	0.057
	x-vector	0.085	0.273	0.255	0.800	1.138	0.214	0.013	0.032
	WavLM + AM	0.228	0.433	0.410	0.839	2.868	0.370	0.022	0.064
StyleCap w/ Llama 2	Mel-spectrogram + AM	0.160	0.358	0.332	0.821	2.109	0.295	0.022	0.066
	x-vector	0.076	0.262	0.239	0.799	1.107	0.213	0.016	0.042
	WavLM + AM	0.273	0.497	0.469	0.855	3.471	0.434	0.023	0.073

Appendix: 実験結果詳細(データ拡張あり)

○評価指標:

□B@4: BLEU4, R: ROUGE, M: METEOR, BS: BERTScore, C: CIDEr-D, S: SPICE scores

Model	Speech encoder	B@4(↑)	R(↑)	M(↑)	BS(↑)	C(↑)	S(↑)	distict-1(↑)	distict-2(↑)
Transformer encoder-decoder	Mel-spectrogram	0.140	0.332	0.303	0.814	1.847	0.239	0.018	0.046
	x-vector	0.071	0.244	0.212	0.792	1.046	0.191	0.012	0.027
	WavLM	0.246	0.464	0.441	0.848	3.172	0.404	0.021	0.059
StyleCap w/ GPT-2	Mel-spectrogram + AM	0.164	0.368	0.334	0.822	2.122	0.294	0.021	0.063
	x-vector	0.068	0.260	0.237	0.798	0.895	0.210	0.013	0.033
	WavLM + AM	0.239	0.470	0.439	0.848	3.056	0.403	0.022	0.068
StyleCap w/ Llama 2	Mel-spectrogram + AM	0.165	0.353	0.327	0.818	2.131	0.279	0.024	0.065
	x-vector	0.084	0.259	0.237	0.769	1.157	0.212	0.014	0.034
	WavLM + AM	0.279	0.507	0.479	0.857	3.594	0.447	0.027	0.079

Appendix: Speech encoder出力によるクラス分類

○スタイルファクター分類

- PromptSpeech の音声には (話者の性別, 音高, 話速, 音量) の4つのスタイルファクターに関するラベルが付与されている
 - 性別は (male/female) の2クラス, その他は (low/mid/high) 3クラス
- 実験で示した学習済みの各 Speech encoder の出力からスタイルファクターを予測する1層の線形層を学習

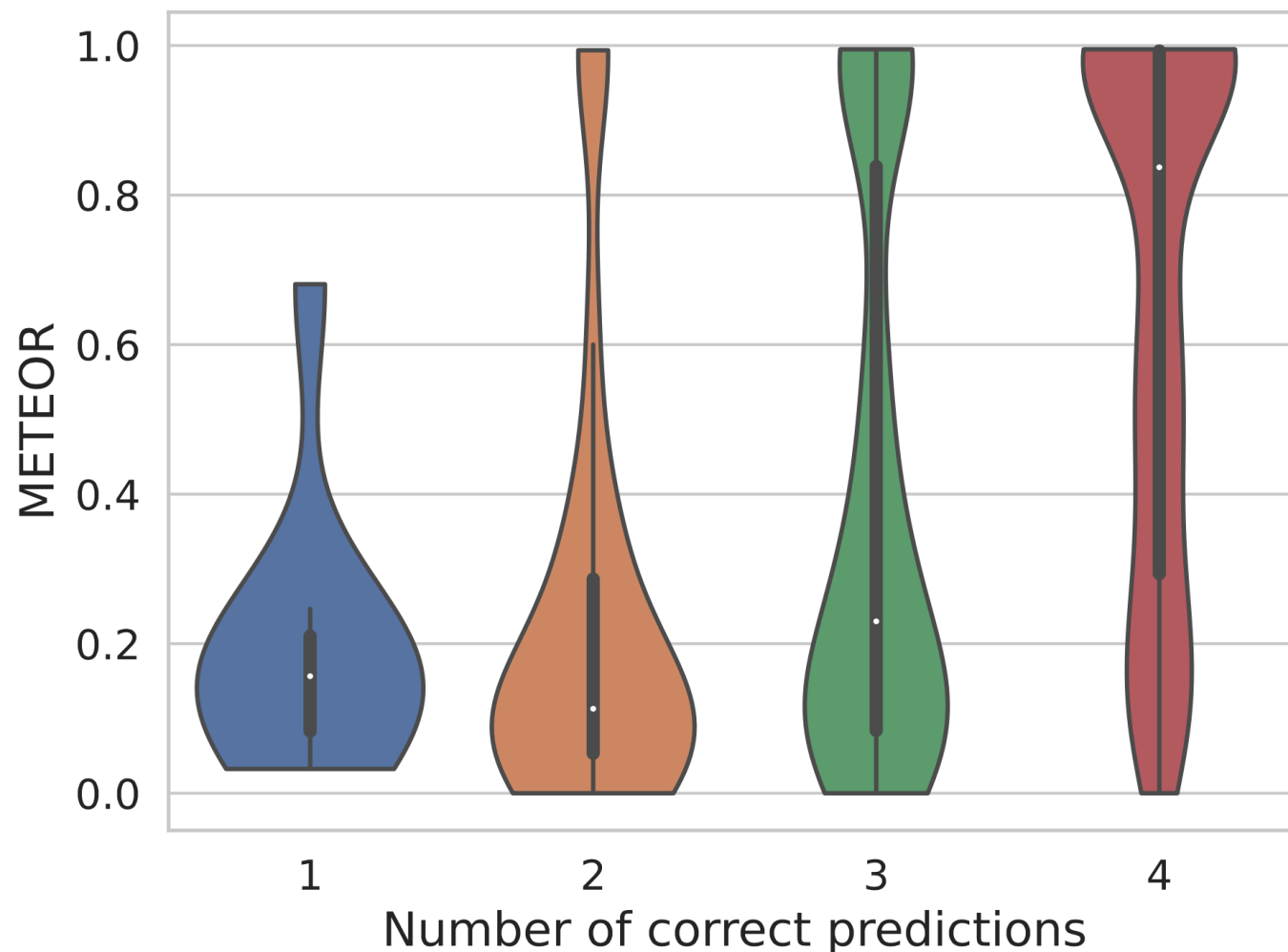
Speech encoder	性別	音高	話速	音量	平均
Mel-spec. + AM	93.8	62.1	67.8	54.7	<u>69.6</u>
x-vec.	94.0	40.5	44.0	49.4	<u>57.0</u>
WavLM + AM	91.0	61.0	85.2	69.9	<u>76.8</u>

分類性能: **WavLM + AM** > others

Appendix: 分類性能とキャプション性能の関係

○スタイルファクター分類の正解数ごとの METEOR 分布

正解数が少なくなるほど
キャプション性能が
劣化している



Appendix: Mapping network の Ablation

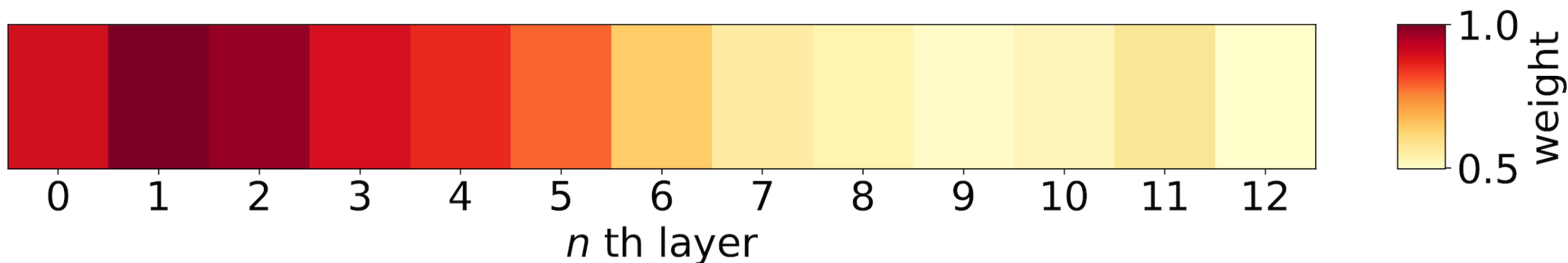
○LLM に入力する Prefix の長さに関する Ablaton study

□モデル: StyleCap w/ WavLM + AM, Llama 2, Sentence rephrasing

Prefix length	1	2	5	10	40	60
METEOR	0.419	0.437	0.468	0.464	0.479	0.459
BERTScore	0.839	0.845	0.853	0.853	0.857	0.852

Appendix: WavLMの隠れ層の分析

○学習済み StyleCap の Speech encoder の WavLM の重み係数



入力に近い層ほど重み係数大きい
→ 入力に近い層ほど発話スタイルに関する情報が豊富

Appendix: LLMへの入力 Prefix の解釈

○Prefix embeddingをコサイン類似度が最も大きい単語に変換

Ground-truth caption	His tone is so high, the volume is very large, and he speak very fast.
Generated Caption (Llama 2)	His sound height is really high, the volume is very large, and he speaks quickly.
Prefix (Llama 2)	角gg元 Dir czas czas czas czas czas czaspper czas czasourceourcecketchetourceource violent czasourceventory czas元clarource wzource spriteventoryppy visual estate gemeins gun cleverggers best western
Generated Caption (GPT-2)	His tone is very high, the volume is very large, and he speak quickly.
Prefix (GPT-2)	heastgaeagementrastructure Dmitentious seeming gaping vertically skyline Explain loudly Explaingae guiding heights technologically soaring firsthandylethrough HIGHstros RIGHT HIGHgaepdffolder Thewhethergae urgently HIGH trave trave traveenged