Mambaを用いた会話の空気感を考慮する音声感情認識技術の開発

Speech-Based Emotion Recognition in Conversation Using Mamba: An Exploration of Emotional Context

地家　康平[*]　加藤　暁浩[*]　長野　紘之[*]　能勢　将樹[*]
Kohei CHIKE　Akihiro KATO　Hiroyuki NAGANO　Masaki NOSE

要　旨

会話における感情認識技術（ERC: Emotion Recognition in Conversation）では，音声に加えて文字起こしテキストや映像といった複数の情報を用いるマルチモーダル手法が有望視されているが，実用面では音声認識誤りの影響や計算コストなどの課題がある．そこで本研究では，音声情報のみに基づいて会話中の感情を認識する新たなアプローチを探求する．具体的には，会話を構成する各発話を効率よく一括処理するためにMambaアーキテクチャを音声感情認識モデルとして応用し，その有効性を評価した．英語データセットIEMOCAPを用いた実験では，会話の一括処理モデルが，発話ごとに逐次処理した場合よりも高い感情認識精度を示した．また，音声のみを入力とした条件でも従来のERCモデルを上回る性能を達成している．さらに本モデルは，会話全体の「空気感」を捉えて感情を判断していることが示唆された．本取り組みは，音声のみを用いてシンプルかつ高精度に会話中の感情を読み取る手法の可能性を示しており，今後の研究や応用展開に向けた重要な一歩となることが期待される．

ABSTRACT

Multimodal approaches to emotion recognition in conversation (ERC) are promising, but automatic speech recognition (ASR) is required to take advantage of the text modality in a practical way, which introduces potential errors and computational overhead. This study explores a speech-only ERC system that employs Mamba to efficiently handle long conversational sequences. Experiments with the IEMOCAP dataset demonstrated improved recognition accuracy compared to a single speech baseline or an existing ERC system limited to speech. Our analysis reveals that while the model does not strongly encode speech order, it does effectively model the overall emotional distribution (i.e., atmosphere) of the conversation. The results highlight the potential of speech-only ERC models, especially those using Mamba, to capture emotion patterns at the conversational level, and suggest directions for future research.

* AIサービス事業本部　AIインテグレーション統括センター
AI Integration Management Center, AI Services Business Division

1. Introduction

Emotion Recognition in conversation (ERC) is a task that tracks emotional states within conversation by considering the context and inter-speaker dependencies[1-3]. While Speech Emotion Recognition (SER) often relies solely on individual utterances, Emotion Recognition in Conversation (ERC) predicts emotions from multiple utterances that make up a conversation. It has attracted attention in recent years due to its potential for applications in a wide range of fields, including customer service[4], health care[5], human-robot interaction[6].

Constructing a superior ERC system requires capturing the temporal dynamics of utterances[2,3]. Existing research has explored modeling long-range contextual information in conversation using recurrent neural networks[7] and Transformers[8], as well as representing dependencies between utterances using graph neural networks[9-11]. Although the mainstream has focused on approaches to textual conversation, multimodal approaches have also been proposed in recent years, achieving good results by combining features of text, video and audio[12-14]. According to ablation studies in 12), 13), text alone produces the highest precision among single modalities, while the combination of text and audio is superior when using two modalities. This suggests that mutual complementation is achieved by extracting linguistic information from text and paralinguistic information from speech.

Despite the promising nature of multimodal approaches, there are several challenges when considering practical applications. The use of text requires automatic speech recognition (ASR), and its misrecognitions can impact performance. Furthermore, high-performance ASR leads to model bloat, and a definitive method for integrating speech and text features is not necessarily known.

On the other hand, speech alone contains a wealth of linguistic and phonological information[15]. Recent self-supervised learning (SSL) models in speech such as Wav2Vec2.0[16], HuBERT[17], WavLM[18], and BEST-RQ[19] achieve state-of-the-art performance on a wide range of speech-related tasks. This suggests the possibility of developing a speech ERC model that relies solely on speech input, directly leveraging the representations learned by SSL models without text from ASR.

This study focuses on building an ERC system using only speech as input. We construct an experimental ERC model that processes conversational temporal dynamics using WavLM features as input and employing Mamba[20,21]. Mamba is an improved state space model (SSM) architecture that enables linear-time sequence processing, allowing it to efficiently handle long-term dependencies in sequences such as conversations.

We conduct experiments on the IEMOCAP dataset: (1) to validate the effectiveness of the model, (2) to analyze how it captures emotions from conversational speech.

For (1), we build a model with a single speech input to compare with SER. The results show that our ERC model achieves higher accuracy than SER (using a single speech input). We also compare the performance of speech-only ERC to existing multimodal models when their input is limited to speech. Our model slightly outperforms them despite its simple architecture.

For (2), We investigate the performance of the model when it is trained on a conversation with randomized utterances. The results show that our model effectively encodes the emotion distribution (i.e., *atmosphere*) of the entire conversation. On the other hand, the order of speech utterances is not strongly encoded. It leads our research interest toward finding a way to encode the relationship between utterances.

Research on speech-only ERC models using Mamba is still limited, and this study demonstrates the potential of models to capture the emotional distribution of conversation from sequences of conversational speech.

2. Model Overview

Fig. 1 illustrates the overall structure of the model used in this study. This model takes a sequence of utterances from a conversation as input and outputs an emotion label for each utterance. It consists of two main components: an SSL feature extractor and a Mamba-based encoder. The feature extractor converts each utterance into frame-level feature embedding vectors. These vectors are then concatenated along the temporal dimension and fed into the Mamba-based encoder. The encoder is expected to process these features and transform them into representations that capture the relationships between individual utterances.

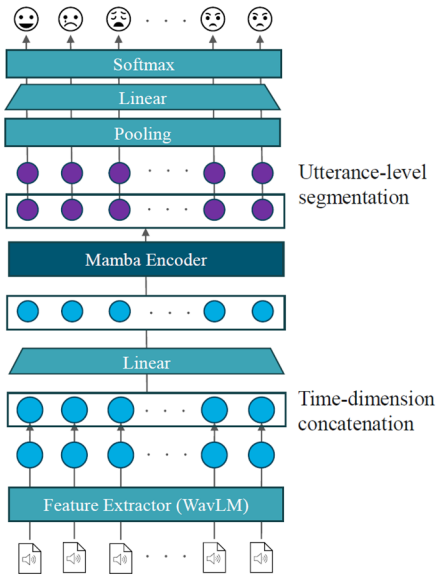

Fig. 1 The overall structure of the model. Each utterance in the conversation is converted into frame-by-frame WavLM features, concatenated in the temporal dimension, and fed to the Mamba Encoder. The encoded features are again divided into utterance units and fed to the latter classifier.

2-1 Task Definition

A conversation, denoted by C, is represented as a sequence of utterance-speaker pairs:

$$C = \{(u_1, s_1), (u_2, s_2), \ldots, (u_n, s_n)\}$$

where u_i represents the i-th utterance and s_i denotes the corresponding speaker.

For each utterance u_i, an emotion label y_i is assigned from a set of possible emotion categories $\mathcal{Y}$ (e.g., happy, sadness, anger, neutral).

The goal is to learn a function

$$f: \{(u_1, s_1), (u_2, s_2), \ldots, (u_n, s_n)\} \rightarrow \{y_1, y_2, \ldots, y_n\}$$

which predicts the optimal emotion label for each utterance using contextual information from the conversation and interactions between speakers.

2-2 Feature Extractor

To extract robust utterance embeddings, we employ the pre-trained and frozen WavLM-large model. This model, a 25-layer transformer trained on 94,000 hours of audio data (LibriLight[22], GigaSpeech[23], and VoxPopuli[24]), has demonstrates strong performance in emotion recognition tasks[18]. We extract features from the 21st layer instead of using the final layer. This approach is known to be effective for emotion recognition task in recent studies[25,26].

2-3 Mamba-based Encoder

The WavLM features for each utterance are concatenated and input into the model as a single sequence.

To handle long-sequence data, we employ Mamba, which enables dynamic inference and maintains efficiency for long sequences compared to Transformers[20]. Fig. 2 shows the structure of the Mamba encoder.

Recent studies have demonstrated the effectiveness of Mamba in various speech processing tasks[27-30]. Our model is based on the *BiMamba* structure from 27). The encoder layer uses bidirectional sequence modeling to process the entire speech signal. Feed-Forward (FF)

module before and after the Mamba blocks to provide nonlinearity. The FF module doubles the embedding dimension, applies Swish[31] as the activation function, and then returns to the original dimension. It also incorporates Layer Norm (LN) and Dropout to stabilize learning and suppress overfitting. The architecture of the FF module is delineated as follows:

$$\mathrm{FF}(x) = \mathrm{Dropout}(W_2\,\tilde{x} + b_2)$$

$$\tilde{x} = \mathrm{Dropout}\big(\mathrm{Swish}(W_1\,\mathrm{LN}(x) + b_1)\big)$$

where

$$x \in \mathbb{R}^{T\times d}$$

$$W_1 \in \mathbb{R}^{d\times 2d}, \quad b_1 \in \mathbb{R}^{2d}$$

$$W_2 \in \mathbb{R}^{2d\times d}, \quad b_2 \in \mathbb{R}^{d}$$

Here, T represents the sequence length (total number of frames), and d is the encoder dimension. W_1, W_2, b_1, b_2 are trainable parameters.

The output, processed by the Mamba layer as a series of long data, is again broken down into utterance units and passed to the linear classifier consisting of the downstream pooling layer and the linear layer.

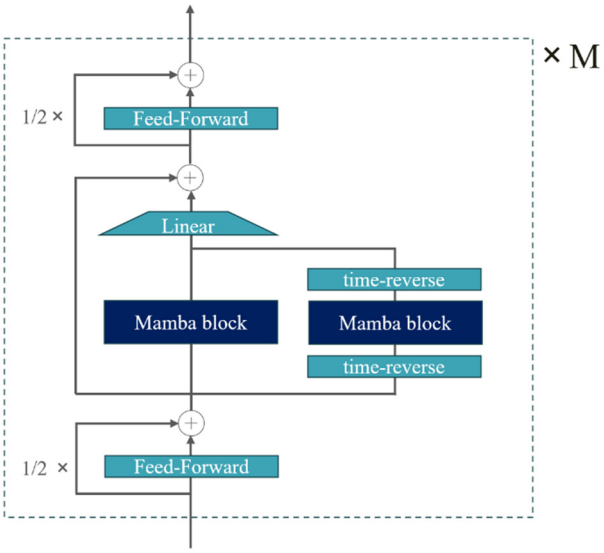

Fig. 2 Mamba encoder. The entire conversation sequence is processed in two Mamba blocks, one forward and one reverse, flanked by Feed-Forward modules.

3. Experiments

3-1 Dataset

We use IEMOCAP[32] dataset. This dataset consists of dialogues between five pairs of male and female speakers, containing a total of 151 conversations. Each utterance is labeled by multiple annotators with 11 different labels.

Four of these labels are significantly fewer than the others. The label *xxx* indicates an utterance for which the annotators did not reach a majority agreement. Table 1 shows the statistics of the IEMOCAP dataset. Previous studies often use subsets with four or six major label types for training. In this study, we use the six main labels (*fru, neu, ang, sad, exc, hap*) as correct response labels. However, to account for the dynamics of the conversation, we use all utterances in the dialog, including *xxx* and fewer labels, as input to the model.

Table 1 Statistics of IEMOCAP dataset. "xxx" identifies samples for which there was no agreement among annotators.

Emotion	# utterances
fru	1849
neu	1708
ang	1103
sad	1084
exc	1041
hap	595
xxx & minor	2659

The dataset consists of five sessions of dialogues between different male-female pairs. Each session contains approximately 30 dialogues. The evaluation is done using leave-one-session-out 5-fold cross-validation.

3-2 Evaluation Metrics

For comparison with previous studies, the model performance is measured using Weighted Accuracy (WA) and Weighted F1 Score (W-F1). WA represents the

accuracy of the model while considering the class distribution, ensuring that classes with different sample sizes are appropriately weighted. Similarly, W-F1 is the weighted version of the F1 score, where the contribution of each class is scaled according to its prevalence in the dataset, providing a more balanced assessment of performance across imbalanced classes.

3-3 Implementation Details

The Mamba encoder has a 256-dimensional input and consists of three layers. Our preliminary experiments have indicated that adding more layers does not yield significant improvements. We use the same hyperparameters of Mamba as the official implementations[*1].

We use cross-entropy loss as the objective function and a single conversation as one batch.

As referred in 3.1, the "xxx" and minority label utterances are input to the model, but the output is not used for training. For the entire dataset, the maximum, minimum, and average number of frames per conversation were 26684, 4402, and 14776.4, respectively.

To compare our model with the SER, we also trained our model with a single speech input. The batch size was set to 128, and all other hyperparameters were the same as when an entire conversation was input.

We used RAdamScheduleFree[*2] as the optimizer. RadamScheduleFree is the Radam[33] version of the Schedule-Free optimizer[34], which does not require warmup, learning rate scheduler, and training end time specifications. We set an initial learning rate to 1.0e-4.

Training was conducted using a single A6000 GPU for 60 epochs, and the model from the epoch with the highest Weighted F1 score was selected.

*1 https://github.com/state-spaces/mamba

*2 https://github.com/facebookresearch/schedule_free?tab=readme-ov-file#releases

4. Results & Discussion

4-1 Comparison of audio-only emotion recognition with prior work

Table 2 presents the performance on IEMOCAP dataset. To compare the performance our ERC model with SER, we present the results of our model trained on single utterance. To compare the performance of ERC, we also present results from the previous study with modality-specific results[12,13].

Table 2 Performance on IEMOCAP dataset. The best result in audio only model is in bold.

Model	WA [%]	W-F1 [%]
Ours	**61.5**	**61.2**
Ours (Single Utterance Input)	57.2	57.0
Ma et al[12]. (Audio Only)	59.8	59.3
Shou et al[13]. (Audio Only)	58.6	58.8
Ma et al[12]. (Text + Audio)	72.5	72.8
Shou et al[13]. (Text + Audio)	71.3	70.2

Our model, which learns and infers using the entire conversation as input, shows improved accuracy from the model trained with one utterance as input. In addition, compared to previous studies that limited the modality to speech, our model shows slightly better results. However, it still falls short of state-of-the-art emotion recognition models that combine text and audio. This may be due to information that cannot be captured solely from speech, and/or the model's insufficient ability to grasp the context of the conversation.

In the following two subsections, we analyze in detail what our model captures.

4-2 Can the model capture the context of the conversation?

To examine whether the model captures the temporal dynamics (i.e., context) of the conversation, we conduct a shuffling experiment on the training data. Specifically, we compare the performance of models trained on IEMOCAP sessions 1-4 as the training data, and session 5 as the test data, under the following four conditions:

- **No Shuffle**: The conversation content, utterance order, and speakers are all fixed.
- **Intra-Conversation Shuffle**: The utterance order within each conversation in the training data is shuffled, while other aspects remain unchanged.
- **Inter-Conversation Shuffle (Speaker Fixed)**: Utterances are randomly shuffled across different conversation in the training data, but the speaker pairs remain fixed.
- **Inter-Conversation Shuffle**: Both utterances and speakers are randomly shuffled across different conversation in the training data.

The conversation content, utterance order, and speakers in the test data are fixed across all conditions. If temporal dynamics such as utterance order are important, shuffling the training data should degrade performance.

Table 3 shows that the No Shuffle condition yielded the highest accuracy. The Inter-Conversation Shuffle condition brings a decrease in accuracy of approximately 3%, similar to the performance when only a single utterance is input. This is likely because the lack of coherence in the conversation content prevents the model from accurately estimating emotions. In contrast, the accuracy decrease in the Intra-Conversation Shuffle condition is limited to about 1.4%, which was less than the decrease in the Inter-Conversation Shuffle condition.

These results suggest that the model used in this study does not consider the utterance order (context) very much but instead judges emotions by capturing the overall emotional distribution, or *atmosphere* of the conversation.

Table 3 Changes in accuracy when altering the order of data within a batch.

Batching Method	WA [%]
(1) No Shuffle	**61.5**
(2) Intra-Conversation Shuffle	60.1
(3) Inter-Conversation Shuffle (Speaker Fixed)	58.2
(4) Inter-Conversation Shuffle	58.6
(5) Single Utterance	57.2

4-3 Can the model feel the *atmosphere* of the conversation?

In natural conversation, the various emotions rarely appear in the same order, and a few emotions often dominate. For example, a conversation involving laughter is likely to contain a lot of *hap* (happiness), while an argument is likely to contain a lot of *ang* (anger) or *fru* (frustration).

To verify this hypothesis, we visualized the distribution of emotion labels in each conversation in IEMOCAP session 5 (Fig. 3). The results show that in 71% (22/31) of the conversation, a specific one or two emotions, excluding *xxx/minor* labels, accounts for more than 50% of the conversation. These dominant emotions can be considered to form the *atmosphere* of the conversation.

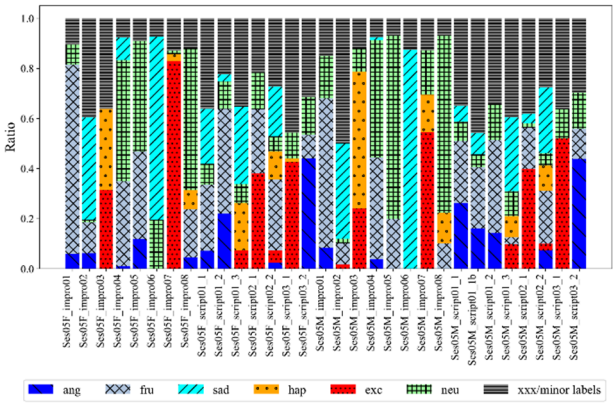

Fig. 3 The distribution of emotions within each dialogue in IEMOCAP Session 5. Horizontal axis represents conversation ID. The vertical axis shows the ratio of utterances of each emotion to the total number of utterances in the conversation.

Table 4 Relative accuracy improvement rate [%] for each emotion compared to single-utterance model. Underlined values indicate the dominant emotion (≒ *atmosphere*) within the conversation. "-" indicates that the specific emotion label does not exist in the conversation.

Conversation ID	ang	exc	fru	hap	neu	sad
Ses05F_impro01	-	-	+34.7	-	-	-
Ses05F_impro06	-	-	-	-	+16.7	+22.6
Ses05F_impro07	-	+5.3	-	−100.0	-	-
Ses05F_impro08	+100.3	-	+33.3	-	+19.3	-
Ses05M_impro01	-	-	+39.9	-	-	-
Ses05M_impro03	-	−10.1	-	+132.9	−40.0	-
Ses05M_impro05	-	-	−20.1	-	+23.9	-
Ses05M_impro06	-	-	-	-	-	+20.0
Ses05M_impro07	-	+59.3	-	+100.0	−28.6	-
Ses05M_impro08	-	-	-	-	+25.1	-

To further support the idea that the model captures such *atmosphere*, we focus on conversation where one emotion is dominant and compared the accuracy between inputting the entire conversation and inputting only a single utterance. As shown in Table 4, the error rate for major emotions improves in all conversation when the entire conversation is input compared to inputting only a single utterance. This suggests that the model can feel the *atmosphere* of the conversation and can recognize dominant emotions more accurately.

On the other hand, the recognition accuracy of minority emotions in the conversation shows decrease in some cases. This is likely a side effect of the model's predictions being pulled towards the dominant emotions within the conversation. In some cases, it is important not to miss emotions that are not the majority in conversation. The results suggest room for further improvement in our model (e.g., developing a stronger encoding of relations between utterances) as a future direction.

5. Conclusion

In this study, we explored the potential of a speech-based Emotion Recognition in conversation (ERC) system using Mamba, a linear-time sequence modeling architecture. Our experiments on the IEMOCAP dataset demonstrated improved accuracy compared to a single speech baseline or an existing ERC system limited to speech. This suggests that the model can effectively capture the overall emotional tone within a conversation, even without relying on textual information or Automatic Speech Recognition (ASR).

However, our analysis also revealed limitations in the model's ability to capture complex temporal dynamics and inter-utterance relationships, indicating a need for further investigation into the methods that explicitly model conversational context. Future research directions include exploring alternative training strategies and architectural modifications to enhance the model's ability to capture richer linguistic information and inter-utterance dependencies. Specifically, employing multi-task learning approaches by incorporating auxiliary tasks such as speech recognition or speaker identification could enrich

the Mamba model with stronger linguistic representations. Furthermore, exploring methods to explicitly encode relationships between utterances, such as incorporating relational inductive biases or contrastive learning techniques, could contribute to a more comprehensive understanding of conversational dynamics.

Despite these limitations, this study provides a valuable foundation for future research into speech-based ERC systems and highlights the promise of Mamba as a computationally efficient alternative for modeling conversational data.

References

1) S. Poria et al.: Emotion recognition in conversation: Research challenges, datasets, and recent advances, *IEEE access*, Vol. 7, pp. 100943–100953 (2019).
2) Y. Shou et al.: A comprehensive survey on multi-modal conversational emotion recognition with deep learning, *arXiv preprint arXiv:2312.05735* (2023).
3) Y. Fu et al.: Emotion recognition in conversations: A survey focusing on context, speaker dependencies, and fusion methods, *Electronics*, Vol. 12, No. 22, p. 4714 (2023).
4) Y. Yurtay et al.: Emotion Recognition on Call Center Voice Data, *Applied Sciences* (2024).
5) B. Subramanian et al.: Digital Twin Model: A Real-Time Emotion Recognition System for Personalized Healthcare, *IEEE Access*, Vol. 10, pp. 81155–81165 (2022).
6) C. Zhu, W. Ahmad: Emotion Recognition from Speech to Improve Human-Robot Interaction, *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/ CyberSciTech)*, pp. 370–375 (2019).
7) N. Majumder et al.: Dialoguernn: An attentive rnn for emotion detection in conversations, *Proc. AAAI conference on artificial intelligence*, pp. 6818–6825 (2019).
8) S. Ghosh et al.: Context and Knowledge Enriched Transformer Framework for Emotion Recognition in Conversations, *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (2021).
9) D. Ghosal et al.: Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, *arXiv preprint arXiv:1908.11540* (2019).
10) Y. Wei et al.: MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video, *Proc. 27th ACM international conference on multimedia*, pp. 1437–1445 (2019).
11) W. Ai et al.: DER-GCN: Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition, *IEEE Transactions on Neural Networks and Learning Systems* (2024).
12) H. Ma et al.: A transformer-based model with self-distillation for multimodal emotion recognition in conversations, *IEEE Transactions on Multimedia* (2023).
13) Y. Shou et al.: Revisiting Multi-modal Emotion Learning with Broad State Space Models and Probability-guidance Fusion, *arXiv preprint arXiv:2404.17858* (2024).
14) X. Li et al.: Mamba-Enhanced Text-Audio-Video Alignment Network for Emotion Recognition in Conversations, *arXiv preprint arXiv:2409.05243* (2024).
15) H. Fujisaki: Information, prosody, and modeling-with emphasis on tonal features of speech, *Speech Prosody 2004, International Conference* (2004).

16) A. Baevski et al.: wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in neural information processing systems*, Vol. 33, pp. 12449–12460 (2020).
17) W.-N. Hsu et al.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units, *IEEE/ACM transactions on audio, speech, and language processing*, Vol. 29, pp. 3451–3460 (2021).
18) S. Chen et al.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing, *IEEE Journal of Selected Topics in Signal Processing*, Vol. 16, No. 6, pp. 1505–1518 (2022).
19) C.-C. Chiu et al.: Self-supervised learning with random-projection quantizer for speech recognition, *International Conference on Machine Learning*, pp. 3915–3924, PMLR (2022).
20) A. Gu, T. Dao: Mamba: Linear-time sequence modeling with selective state spaces, *arXiv preprint arXiv:2312.00752* (2023).
21) T. Dao, A. Gu: Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality, *arXiv preprint arXiv:2405.21060* (2024).
22) J. Kahn et al.: Libri-light: A benchmark for asr with limited or no supervision, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673 IEEE (2020).
23) G. Chen et al.: Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio, *arXiv preprint arXiv:2106.06909* (2021).
24) C. Wang: VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation, *Proc. 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003 (2021).
25) Z. Zhu, Y. Sato: Deep investigation of intermediate representations in self-supervised learning models for speech emotion recognition, *2023 IEEE International conference on acoustics, speech, and signal processing workshops (ICASSPW)*, pp. 1–5, IEEE (2023).
26) S.-w. Yang et al.: A Large-Scale Evaluation of Speech Foundation Models, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).
27) X. Zhang et al.: Mamba in Speech: Towards an Alternative to Self-Attention, *arXiv preprint arXiv:2405.12609* (2024).
28) K. Miyazaki, Y. Masuyama, M. Murata: Exploring the capability of mamba in speech applications, *arXiv preprint arXiv:2406.16808* (2024).
29) J. Lin, H. Hu: Audio Mamba: Pretrained Audio State Space Model For Audio Tagging, *arXiv preprint arXiv:2405.13636* (2024).
30) A. Plaquet et al.: Mamba-based Segmentation Model for Speaker Diarization, *arXiv preprint arXiv:2410.06459* (2024).
31) P. Ramachandran et al.: Searching for Activation Functions, *CoRR*, Vol. abs/1710.05941 (2017).
32) C. Busso et al.: IEMOCAP: Interactive emotional dyadic motion capture database, *Language resources and evaluation*, Vol. 42, pp. 335–359 (2008).
33) L. Liu et al.: On the variance of the adaptive learning rate and beyond, *arXiv preprint arXiv:1908.03265* (2019).
34) A. Defazio et al.: The road less scheduled, *arXiv preprint arXiv:2405.15682* (2024).