Cutting Through Overload: Efficient Token Dropping for Speech Emotion Recognition in Multimodal Large Language Models

Jaime Bellver-Soler, Mario Rodríguez-Cantelar, Ricardo Córdoba, Luis Fernando D'Haro

Speech Technology and Machine Learning Group (THAU) - ETSI de Telecomunicación
Av. Complutense 30, 28040, Madrid, Spain - Universidad Politécnica de Madrid
{jaime.bellver, mario.rcantelar, ricardo.cordoba, luisfernando.dharo}@upm.es

Correspondence: jaime.bellver@upm.es

Abstract

Recent developments in Multimodal Large Language Models (MLLMs) have provided novel insights into Speech Emotion Recognition (SER). However, combining high-dimensional speech signals with textual tokens can lead to a rapid growth in input tokens, increasing computational costs and inference times. This "token overload" also risks shadowing essential textual cues, affecting the reasoning capabilities of the language model and diluting emotional information crucial to accurate SER.

In this paper, we explore different token drop methods that mitigate excessive token counts while preserving both emotional nuances and the core linguistic capabilities of the model. Specifically, we compare various efficient pooling approaches to produce a compact representation. Our preliminary findings suggest that these techniques can reduce computational costs without decreasing SER accuracy.

1 Introduction

Speech Emotion Recognition (SER) has garnered growing interest due to its potential in various applications, including human-computer interaction, mental healthcare, and education. Although single-modality methods, such as text-based emotion analysis or audio emotion recognition, have proven effective (Maruf et al., 2024; George and Ilyas, 2024), emotional data in real-world scenarios often integrate multiple modalities. This has led to increased interest in the use of Multimodal Large Language Models (MLLMs) to exploit knowledge from different data sources and improve emotional reasoning (Chandraumakantham et al., 2024).

Recent advances in MLLMs have demonstrated remarkable performance in audio analysis (Chu et al., 2023). However, there are challenges to applying MLLMs to SER. One of the key obstacles is the rapid increase in the number of multimodal tokens, which drastically expands the size of the input of the model. These multimodal token embeddings can increase computational costs (Ju et al., 2023), prolong inference times, and potentially shade text tokens during the model's attention process (Zhang et al., 2024), thus reducing overall performance.

To address these limitations, researchers have begun exploring token drop strategies (Li et al., 2023a; Zhang et al., 2023b; Rekesh et al., 2023; Gaido et al., 2021; Li et al., 2023b; Yao et al., 2024; Fathullah et al., 2023; Liu et al., 2024; Arif et al., 2024), with the aim of ensuring a more balanced and efficient integration of audio and textual information within MLLMs.

This article builds upon these efforts by recognizing that some existing approaches in the literature may become complex due to the large number of parameters or the complexity of the training. In this work, we explore simple pooling methods that help to control the excessive growth of acoustic tokens. By reducing the token overload on the language model, we can preserve its core linguistic capabilities while enhancing its ability to recognize audio-based emotions, crucial for dialogue systems that must handle both textual and emotional cues effectively. We evaluated how these pooling strategies affect computational costs, inference speed, and prediction accuracy, showing new insights into optimizing MLLMs for SER, improving dialogue systems, and enhancing human-computer interactions.

2 Related Work

SER has evolved with the appearance of multimodal approaches and Large Language Models (LLMs) that incorporate audio inputs. Early research often focused on single-modality solutions, either through acoustic features or text-based analysis, to detect emotions. However, these methods struggled to generalize in different contexts and linguistic styles, motivating the development of

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 284–289
May 27–30, 2025.

multimodal systems that merge information from speech, text, and sometimes visual cues (Lian et al., 2023). Recent work has shown that integrating text and audio using MLLMs can produce more robust and nuanced emotion predictions (Deshmukh et al., 2024; Tang et al., 2024a).

Afterwards, a series of MLLMs have emerged, extending the capabilities of LLMs to handle different input types (Chu et al., 2023; Tang et al., 2024b). These architectures have shown promising results on tasks ranging from automatic speech recognition to generic audio understanding (Gong et al., 2023; Zhang et al., 2023a). However, many existing MLLMs either rely on supervised training of additional transformer modules or require extensive fine-tuning on specific downstream tasks, making them computationally expensive and less flexible for broader SER applications.

Despite progress in audio-based LLMs, a key challenge remains: the rapid increase in token counts (token overload) when merging high-dimensional audio and textual representations. Token overload can degrade model performance, increase computational costs, and slow down inference, problems especially salient in real-time or large-scale deployments (Li et al., 2023b; Shang et al., 2024). To handle this, a variety of token dropping methods have been proposed. Simple statistical techniques, such as mean pooling, can compress feature representations at minimal cost. Also, concatenation-based strategies can combine tokens in pairs or in groups to reduce the sequence length (Fathullah et al., 2023). More complex methods employ n-dimensional convolutions (Zhang et al., 2023b), lightweight Q-Formers (Li et al., 2023a), or architectures such as Fast Conformer (Rekesh et al., 2023) and Connectionist Temporal Classification (CTC) (Gaido et al., 2021) to remove redundant information. Further advancements have been explored in vision frameworks, such as Liu et al. (2024); Arif et al. (2024); Shang et al. (2024) that dynamically prune tokens based on attention scores or local content similarity. Although originally proposed for image or video tasks, these strategies offer valuable insights for audio-based MLLMs.

In this paper, we focus on evaluating simple and efficient token dropping methods for SER tasks using MLLMs. Rather than relying on complex architectures or parameter-heavy models, we explore straightforward pooling techniques to optimize the token efficiency of multimodal inputs. Our approach aims to reduce computational costs and inference times while maintaining the ability of the model to capture emotional nuances. By integrating simplified token reduction modules into a multilingual SER pipeline, we demonstrate that efficient length adaptation techniques can achieve competitive performance.

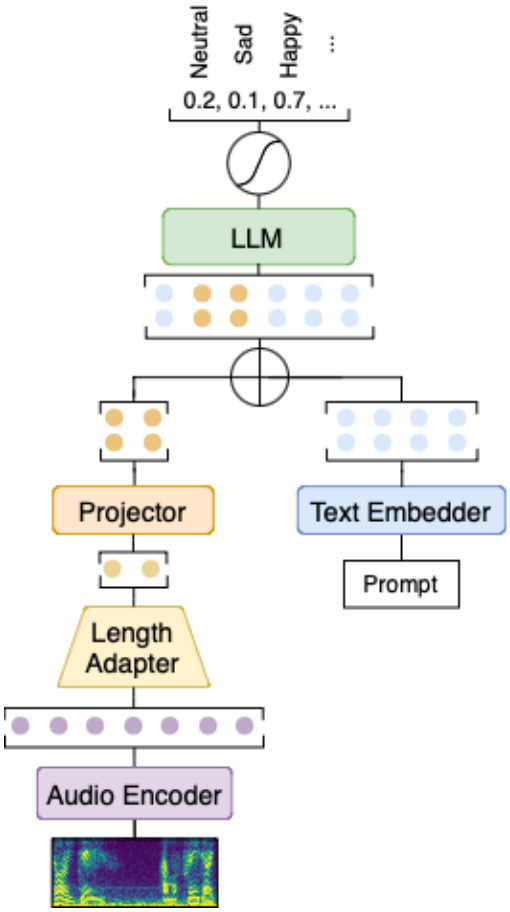

Figure 1: MLLM architecture for SER with a length adapter.

3 Methods

To address the challenges of SER within multilingual and multimodal contexts, we propose a methodology that integrates high-dimensional speech and text signals into a unified framework. Our approach combines an audio encoder based on transformers, a linear projection layer, and an LLM, creating a multimodal architecture (see Figure 1 for a detailed diagram of the model architecture).

We employ the *Whisper-large-v3* encoder (Radford et al., 2022), a state-of-the-art model known for its ability to extract rich phonetic features from audio log-mel spectrograms (Gong et al., 2023; Zhang et al., 2023a). The encoded audio representations are then processed through a linear projec-

tor, which changes their dimensionality to align with the embedding space of the LLM (Chu et al., 2023). For the text component, we use *Gemma-2-2B-it* (et al., 2024), an LLM designed to handle diverse linguistic contexts and capable of reasoning over multilingual inputs.

We address the token overload challenge by incorporating length adaptation strategies that compress high-dimensional audio embeddings into more compact representations. These strategies range from simple statistical pooling methods, such as *Mean* pooling, to more complex approaches like *Convolutional* (*Conv*) compression (Zhang et al., 2023b), Concatenation (*Concat*) (Fathullah et al., 2023), and attention mechanisms (Vaswani et al., 2023).

We begin with *Mean* pooling, which aggregates embeddings by straightforward averaging. We then employ *Conv*, using convolutional filters to extract salient features, and a *Concat* approach that pairs tokens, effectively halving the sequence length while doubling the dimensionality.

We explore attention-based methods starting with *Attn-Mean*, which averages the output of the attention layer, and *Attn-Q-Mean*, which introduces a global query vector. Specifically, instead of deriving the query from each input token, we first perform a mean pooling across the entire sequence of input embeddings $X \in \mathbb{R}^{L \times D}$ (see Equation 1).

$$Q = \Big(\frac{1}{L}\sum_{i=1}^{L} x_i\Big) W_q, \tag{1}$$

where x_i is the $i-th$ token embedding in X, and $W_q \in \mathbb{R}^{D \times d_k}$ projects the averaged embedding into the query space. The keys K and values V are computed using standard linear projections from X. The final compact representation is obtained via a standard scaled dot product attention mechanism that uses Q, K, and V (Vaswani et al., 2023).

To evaluate our framework, we first develop text-only baseline models. We employ transcriptions generated with *Whisper-large-v3* from speech input, and we use a frozen *Gemma-2-2b-it* to predict emotions only based on textual information. Building on these baselines, we integrate audio features into the pipeline, in which audio embeddings are combined with text tokens. We train only the linear projector and the length adapter layers, ensuring that the LLM retains its original capabilities.

Performance was measured using the F1 macro score and Weighted Accuracy (WA) to account for class imbalances. For each evaluation, we deployed a 5-fold cross-validation strategy and report the mean F1 macro and WA, along with their standard deviations.

We emphasize multilingual SER, using datasets in Spanish, German, and French to validate the generalization of our approach using datasets from MEACorpus, EmoDB, and Oreau (Pan et al., 2024; Burkhardt et al., 2005; Kerkeni et al., 2020). The three datasets contain emotion labels for Fear, Sad, Happy, Angry, Disgust, and Neutral, EmoDB also includes Boredom, while Oreau Surprise. The MEACorpus dataset suffers from imbalanced class distributions, which present additional challenges for robust modeling, although both EmoDB and Oreau have more balanced class distributions. Furthermore, the data in MEACorpus are derived from natural YouTube videos, reflecting real-world, spontaneous emotions, while EmoDB and Oreau datasets consist of acted recordings, which provide more controlled but less naturalistic emotional expressions. Our design also prioritizes computational efficiency, enabling faster inference times without compromising accuracy, an essential factor for the deployment of real-world dialogue systems.

4 Experiments

Our preliminary experiments focus on selecting the optimal components for the MLLM architecture, tuning hyperparameters [1], and refining a prompt for the LLM [2]. *Whisper-large-v3* was selected as audio encoder, while a linear projector was chosen for its effectiveness (Chu et al., 2023) and simplicity in aligning audio embeddings with the input requirements of the LLM. *Gemma-2-2B-it* was chosen as the LLM due to its remarkable performance in handling multimodal inputs and reasoning in both text and audio (et al., 2024).

To establish SER baselines, we first evaluated a text-only model, where the LLM remained frozen and predictions were made solely from the transcriptions of the speech input. This text-only baseline achieved an average F1 macro score of 0.23 and a WA of 0.29 across the three datasets.

The integration of audio and text modalities was evaluated through MLLMs, testing variations in length adaptation strategies. Detailed results for each dataset can be found in Table 1, while the overall averages are summarized in Table 2. In the ini-

[1]The hyperparameters used can be found in Annex A.

[2]The prompt used can be found in Annex B.

Adapter	MEACorpus (ES)		EmoDB (DE)		Oreau (FR)	
	WA	F1 macro	WA	F1 macro	WA	F1 macro
None	0.72 ± 0.01	0.62 ± 0.04	0.39 ± 0.05	0.36 ± 0.06	0.69 ± 0.07	0.69 ± 0.07
Mean	0.74 ± 0.01	0.64 ± 0.03	$\mathbf{0.60} \pm 0.06$	$\mathbf{0.53} \pm 0.05$	0.79 ± 0.04	0.78 ± 0.05
Concat	0.72 ± 0.01	0.57 ± 0.02	0.41 ± 0.02	0.40 ± 0.03	$\mathbf{0.92} \pm 0.32$	0.73 ± 0.02
Conv	0.73 ± 0.02	0.63 ± 0.06	0.47 ± 0.05	0.42 ± 0.07	0.84 ± 0.03	$\mathbf{0.84} \pm 0.02$
Atten-Mean	$\mathbf{0.76} \pm 0.01$	$\mathbf{0.69} \pm 0.04$	0.55 ± 0.08	0.50 ± 0.07	0.83 ± 0.03	0.82 ± 0.03
Attn-Q-Mean	0.75 ± 0.01	0.67 ± 0.03	0.53 ± 0.02	0.47 ± 0.02	0.82 ± 0.05	0.81 ± 0.06

Table 1: Average WA and F1 macro scores across the 5-folds and its standard deviation are presented in columns under each dataset. The row labeled "*None*" corresponds to the model without a length adapter.

Adapter	Trainable Params	Speed-up	Acoustic Tokens	Mean WA	Mean F1 macro
None	0	0% -	170 -	0.60 ± 0.03	0.56 ± 0.06
Mean	$1.2M$	21% ↑	1 ↓↓	0.70 ± 0.02	0.65 ± 0.04
Concat	$1.8M$	16% ↑	85 ↓	0.68 ± 0.02	0.56 ± 0.02
Conv	$1.5M$	3% ↑	85 ↓	0.68 ± 0.03	0.63 ± 0.05
Atten-Mean	$1.6M$	22% ↑	1 ↓↓	$\mathbf{0.71} \pm 0.03$	$\mathbf{0.67} \pm 0.05$
Attn-Q-Mean	$1.7M$	**26%** ↑↑	1 ↓↓	0.70 ± 0.02	0.65 ± 0.03

Table 2: WA and F1 macro averaged across datasets, along with the trainable parameters (Trainable Params), the decrease of inference time (Speed-up) with respect to the alternative without length adapter (which achieves 18 iterations per second on a single A100 GPU), and the number of acoustic tokens.

tial configuration, labeled "*None*", the projected audio embeddings were directly fed into the language model without any token dropping. While this approach preserved the complete acoustic fidelity, it also introduced a token overload, resulting in an average of 170 acoustic tokens per sample across the three datasets. Although it achieved a mean WA of 0.60 and an F1 macro of 0.56, exceeding the text only baseline, this increased token count substantially increased the computational cost, with inference times up to 26% higher compared to the text only model.

To address this, we implemented various length adaptation techniques to compress high-dimensional audio embeddings. First, simple pooling methods, such as *Mean* pooling, improved performance to a mean WA of 0.70 and an F1 macro of 0.65. Next, *Conv* and *Concat* both achieved WA scores of 0.68, with macro F1 scores of 0.63 and 0.56, respectively. Finally, attention-based approaches (*Attn-Mean* and *Attn-Q-Mean*) further boosted overall performance. Attn-Mean achieved the highest metrics, while *Attn-Q-Mean* also performed strongly.

Table 2 also details computational trade-offs, including inference speed-ups relative to the no-adapter baseline (*None*). *Mean*, *Attn-Mean*, and *Attn-Q-Mean* all compress the acoustic representation to a single token, achieving speed-ups of 21%, 22%, and 26%, respectively. In contrast, *Concat* and *Conv* halve the number of tokens, resulting in speed-ups of 16% and 3%. In particular, Attn-Mean strikes an optimal balance between accuracy and efficiency, securing the highest F1 macro and WA scores while still offering a 22% speed-up.

5 Conclusion

Our experiments confirm that integrating audio and text in MLLMs significantly enhances SER, surpassing text-only approaches in both WA and F1 macro metrics. However, directly merging audio embeddings can lead to "token overload", increasing computational demands and slowing down inference. By incorporating simple length adapters, we achieved significant inference speed-ups (from 22% to 26%) compared to the baseline, while retaining or improving SER accuracy. Notably, we only fine-tuned a lightweight projector layer, thereby preserving the language reasoning capabilities of the LLM.

In future work, we will explore more advanced token compression strategies and extend our experiments to a broader range of tasks and datasets, aiming for higher scalability and robust performance across diverse multimodal dialogue systems.

Acknowledgements

This work is supported by the European Commission through Project ASTOUND (101071191 — HORIZON EIC-2021-PATHFINDERCHALLENGES-01), by project BEWORD (PID2021-126061OB-C43) funded by MCIN/AEI/10.13039/501100011033 and, as appropriate, by "ERDF A way of making Europe", by the "European Union", and by project INNOVATRAD-CM (PHS-2024/PH-HUM-52) from Comunidad de Madrid.

References

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S. Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. 2024. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. *Preprint*, arXiv:2408.10945.

Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of german emotional speech. In *Interspeech*.

Omkumar Chandraumakantham, N. Gowtham, Mohammed Zakariah, and Abdulaziz Almazyad. 2024. Multimodal emotion recognition using feature fusion: An llm-based approach. *IEEE Access*, 12:108052–108071.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2024. Pengi: An audio language model for audio tasks. *Preprint*, arXiv:2305.11834.

Gemma Team et al. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2023. Prompting large language models with speech recognition abilities. *Preprint*, arXiv:2307.11795.

Marco Gaido, Mauro Cettolo, Matteo Negri, and Marco Turchi. 2021. Ctc-based compression for direct speech translation. *Preprint*, arXiv:2102.01578.

Swapna Mol George and P. Muhamed Ilyas. 2024. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568:127015.

Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass. 2023. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers.

Chen Ju, Haicheng Wang, Zeqian Li, Xu Chen, Zhonghua Zhai, Weilin Huang, and Shuai Xiao. 2023. Turbo: Informativity-driven acceleration plug-in for vision-language models. *Preprint*, arXiv:2312.07408.

Leila Kerkeni, Catherine Cleder, Youssef Serrestou, and Kosai Raoof. 2020. French emotional speech database - oréau.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023b. Llama-vid: An image is worth 2 tokens in large language models. *Preprint*, arXiv:2311.17043.

Hui Lian, Chen Lu, Shengyuan Li, Yunzhi Zhao, Chenglong Tang, and Yu Zong. 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10).

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. 2024. Multi-stage vision token dropping: Towards efficient multimodal large language model. *Preprint*, arXiv:2411.10803.

Abdullah Al Maruf, Fahima Khanam, Md. Mahmudul Haque, Zakaria Masud Jiyad, M. F. Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access*, 12:18416–18450.

Ronghao Pan, José Antonio García-Díaz, Miguel Ángel Rodríguez-García, and Rafel Valencia-García. 2024. Spanish meacorpus 2023: A multimodal speech–text corpus for emotion analysis in spanish from natural environments. *Computer Standards Interfaces*, 90:103856.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. *Preprint*, arXiv:2305.05084.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *Preprint*, arXiv:2403.15388.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024a. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024b. Salmonn: Towards generic hearing abilities for large language models. *Preprint*, arXiv:2310.13289.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. 2024. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *Preprint*, arXiv:2405.20985.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *Preprint*, arXiv:2305.11000.

Hao Zhang, Nianwen Si, Yaqi Chen, Wenlin Zhang, Xukui Yang, Dan Qu, and Xiaolin Jiao. 2023b. Tuning large language model for end-to-end speech translation. *Preprint*, arXiv:2310.02050.

Yi-Kai Zhang, Shiyin Lu, Yang Li, YanQing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. 2024. Wings: Learning multimodal LLMs without text-only forgetting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

A Hyperparameters of MLLM

The hyperparameters of MLLM are shown in Table 3.

Hyperparameter	Value
Learning rate	$10e-4$
Batch size	4
Accumulate gradients	2
Epochs	20
Betas	$(0.9, 0.98)$
Eps	$1e-5$
Weight decay	0
Attention hidden dimension	1280
Attention heads	1
Linear projector dimensions	$[1280, 2304]$

Table 3: Table of hyperparameters used in the MLLM training.

B Prompt for the MLLM

"user: Transcription: {transcription} \n
Audio: {audio} \n
What is the emotion of the speaker?
The possible emotions are: {emotions}. \n
assistanSt: The emotion of the audio is:"