
Project Group 12 Final Report

Lainey Liu

Bowen Ding

Aria Wu

Abstract

Analyzing speech using language models remains a significant challenge in the field of deep learning research, particularly when addressing speaker verification. Current state-of-the-art models struggle to recognize the same speaker across various emotional states and speech patterns, as multiple influencing factors, such as sentiment variations, can obscure speaker identities. Our research builds upon the SALMONN-Lite model framework, pursuing two primary objectives: first, to achieve comparable performance by training our model based on smaller datasets and reduced computational resources, making training LM more accessible; and second, to enhance the verification transparency by providing detailed explanations of how speaker verification decisions are made based on audio traits identified by the language models. In addition, we explored architectural modifications and future fine-tuning techniques to potentially improve verification accuracy beyond the current model. This project contributes to the development of more efficient, explainable, and robust speaker verification systems capable of maintaining performance in diverse speaking contexts.

1 Introduction

Speaker Verification, the technique to decide whether a voice sample truly belong to a claimed entity, matters far beyond its obvious appeal. Unlike fingerprints or face IDs, a microphone is already embedded nearly every phone and smart vehicles, which makes it the only biometric authentication when people are busy with driving, cooking, or other chores. Convolution neural networks (CNNs) based models have achieved a remarkable success for automatic speaker verification. X-vector employed a 5-layer TDNN over a 1-D time axis, which is a first to adopt such architecture to this problem [4]. Subsequently, there is ECAPA-TDNN combining the TDNN and SE-Res2Net bottleneck with channel attention, which made even bigger breakthrough with a less than 1% error rate [6].

Despite the amazing progresses, CNNs still has its limitations as it lacks the global context fusion. To overcome such drawbacks, researchers tried to introduce Transformers into the problem. Started by UniSpeech and WavLM, though they were powerful, they suffer from the complicated pre-train steps and huge amounts of parameters. MFA-Conformer tried to employ the Convolution-augmented Transformer (Conformer) with an easy-to-implement and effective backbone for speaker embedding extraction provided a even more robust performance [22]. Recently, a model named SALMONN builds on top of the prior works and give a single LLM “generic hearing” abilities by glueing together three frozen building blocks, plus two light trainable adapters [20].

Inspired by these recent progresses, we proposed SALMONN-Lite, a novel framework that synergizes speaker embeddings with large language models (LLMs) through an adaptive Q-Former architecture. Unlike prior works either treat verification as a binary classification task or require heavy memory and computing resources, our model introduces three key innovations: (1) A lightweight BERT-based Query Former that dynamically fuses speaker embeddings from ECAPA-TDNN through cross-attention mechanisms [6]; (2) A frozen GPT-2 LLM that generates the verification decisions along with the natural language analysis of the speakers’ voice traits like nasality, breathiness, and roughness, etc. (3) A modality-aligned projection that maps acoustic features to LLM semantic space

without expensive pretraining. Unlike SaLMONN our LLM takes two audio samples as an input to the system and outputs their corresponding differences in terms of voice quality and personality type.

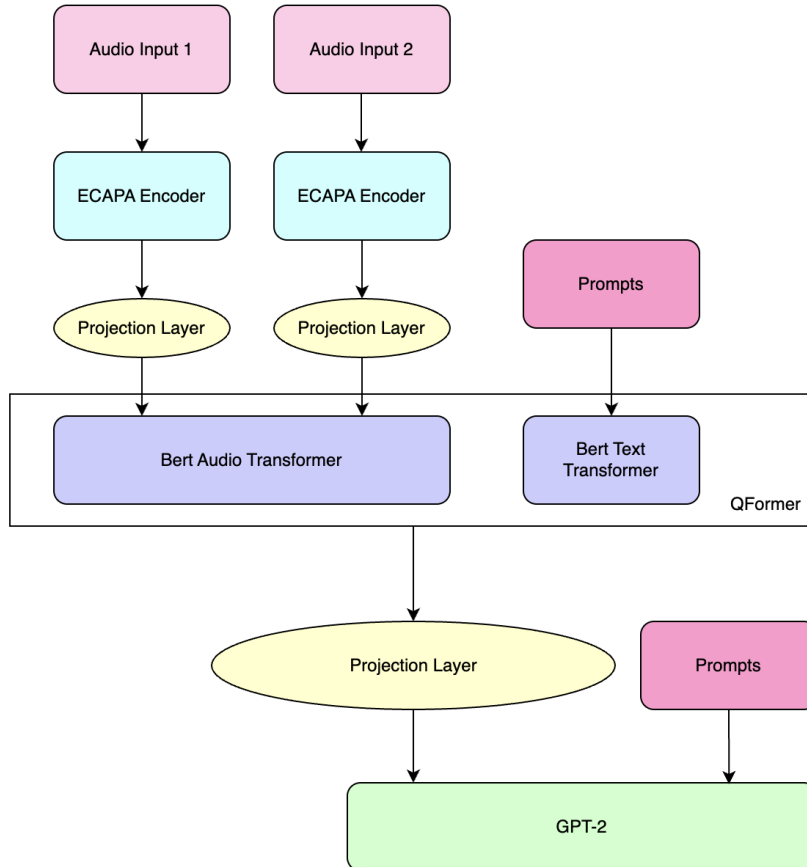
2 Literature Review

Recent work in the speech recognition field has mainly focused on leveraging speech processing by interacting with large-language models (LLMs). This new model framework, often called Audio Language Models (ALMs), has been proven to have great potential. For example, recent research by OpenAI in 2022 published Whisper model that yields an encoder that can handle generic speech translation without task-specific tuning [21]. Despite the great performance of these models at analyzing recognizing audios, few researches has been incorporating the ALM upon speech verification, which the model analyzed if a pair of audio come from the same speaker.

Speaker verification can be very influential with a variety of usages. For example, it can enhance software security by verifying user through their voices. This field is only going to become more popular as there are models like Deepfake audio out there trying to spread false information [23]. Current ALMs seem to focus on understanding one audio file. For example, the SALMONN model [20] focused on automatic speech recognition (ASR) and speech translation by giving a user prompt. The success of SALMONN model and it's key model architecture of using Q-former inspired us to usage a similar architecture while training our model, SALMONN-Lite.

With the potential impact on the speech verification field, our goal lies in developing a model to achieve this. high verification performance while generating a better understanding of speaker's voice traits by leveraging LLMs.

3 Model Architecture



3.1 Dual Auditory Encoders

We adopt a dual-speaker embedding framework for auditory representation. Specifically, we use a pre-trained ECAPA-TDNN model from SpeechBrain as a frozen speaker encoder to extract fixed-dimensional embeddings for two audio inputs. This speaker encoder has demonstrated high accuracy on speaker verification benchmarks and encodes salient voice characteristics such as timbre, pitch, prosody, and speaking rate into fixed 192-dimensional vectors.

Given two audio samples X_1 and X_2 , we compute speaker embeddings:

$$E_1 = \text{ECAPA}(X_1), E_2 = \text{ECAPA}(X_2)$$

where $E_1, E_2 \in \mathbb{R}^{192}$ are the extracted speaker embeddings. These embeddings are then stacked and linearly projected before processing by our lightweight audio transformer.

$$\text{Stack}(E_1, E_2) \in \mathbb{R}^{B \times 2 \times 192} \xrightarrow{\text{Linear Projection}} \mathbb{R}^{B \times 2 \times 768}$$

3.2 Variable-Length QFormer

The Q-Former module in our architecture serves as a bridge between multimodal input features and the language model by extracting relevant semantic representations using a trainable query mechanism. It comprises two transformer submodules: an **audio transformer** and a **text transformer**, both based on the BERT architecture but configured with fewer layers for efficiency.

A fixed number of learnable query tokens $Q \in \mathbb{R}^{(N \times d)}$ are initialized and updated during training. These tokens first undergo self-attention to contextualize across one another, capturing latent dimensions of comparison and speaker relations. The resulting query representations then interact with the encoded features from both audio and text branches via multi-head cross-attention.

Specifically, the audio transformer processes a pair of speaker embeddings (already projected to the hidden size d), producing a contextual audio representation $Z_{audio} \in \mathbb{R}^{2 \times d}$. In parallel, the text transformer processes tokenized prompt input, yielding a sequence of text features $Z_{text} \in \mathbb{R}^{T \times d}$. These are concatenated to form a joint representation:

$$Z = [Z_{audio} ; Z_{text}] \in \mathbb{R}^{(2+T) \times d}$$

The learnable queries attend to this combined sequence through cross-attention, enabling them to selectively aggregate information across modalities. Unlike our baseline model SALMONN using a fix-sized window level QFormer, our implementation flexibly attends over two embeddings and an arbitrary-length prompt, which allows it to preserve the full fidelity of speaker characteristics while avoiding the need for artificial windowing or token masking. The output is a set of enriched latent tokens that serve as a generative prefix for the downstream language model.

3.3 LLM

We employ a pre-trained GPT-2 model as the language model in our architecture. A projection layer is trained to align the output of the Q-Former with the GPT-2 input embedding space, enabling smooth integration without modifying the LLM parameters.

This design ensures that generative outputs are influenced by speaker information and prompt context solely through the learned prefix embeddings, while the language model maintains its original capabilities. Unlike parameter-efficient fine-tuning methods such as LoRA [20], which introduce trainable low-rank adapters within the LLM, our method avoids modifying internal attention layers or weights of GPT-2.

3.4 Training

Our model is trained using a single-stage next-token prediction objective, in which only the Q-Former module, part of the GPT-2 weights and the GPT-2 projection layer are trainable, while the speaker encoder (ECAPA-TDNN) remains entirely frozen. We tried both the GPT-2 weights remain entirely frozen during training and also the learnable scenarios, and it turns out the model has a better performance when the GPT2 weights are learnable as it can better adapt to the speaker comparison

task. The standard prefix-tuning paradigm is also employed where conditioning occurs via external learned tokens rather than updating the model itself. The objective is to generate text conditioned on two input speaker embeddings and a prompt, using cross-attention over audio-text representations.

The model is optimized using a standard cross-entropy loss over the target tokens. Specifically, for a batch of input samples, let $P \in \mathbb{R}^{Q \times d}$ denote the learned prefix embeddings generated by the Q-Former, and $T \in \mathbb{R}^{L \times d}$ the token embeddings from the prompt text. These are concatenated and passed to the GPT-2 decoder:

$$\hat{Y} = \text{GPT-2}([P; T])$$

The training objective is to minimize:

$$\mathcal{L} = - \sum_{i=1}^N \sum_{j=1}^L \log p_{\theta}(y_{i,j} \mid y_{i,<j}, \text{prefix}_i)$$

where $y_{i,j}$ is the j th token of the i th target sequence, and prefix_i is generated from the corresponding speaker pair and prompt via the Q-Former. Padding tokens are ignored during loss computation.

Unlike SALMONN, which uses multi-stage training (pre-training, instruction tuning, and activation tuning) and parameter-efficient tuning (LoRA), our model is trained end-to-end from scratch in a single stage on a curated dataset of speaker pairs and prompts. This setup avoids overfitting to deterministic tasks like ASR or captioning by incorporating diverse prompt formulations and speaker comparisons in training, promoting generalization to more open-ended generation tasks such as speaker reasoning and description.

4 Dataset

We utilize the VoxCeleb-Exp dataset, an extended version of VoxCeleb, as the primary source for constructing speaker comparison tasks. VoxCeleb-Exp contains high-quality, time-aligned speech segments sampled from interviews, talk shows, and public speeches across a wide demographic, offering rich diversity in speaker identity, language background, age, and vocal expression [8].

Each training sample is constructed from a pair of speech recordings: (X_1, X_2) , selected either from the same or different speakers. We apply a pre-trained ECAPA-TDNN speaker encoder to extract fixed-dimensional embeddings that serve as compact representations of speaker characteristics [6]. These embeddings are used as inputs to our Q-Former module, which conditions the generation of text describing speaker similarity and vocal traits.

To promote high-level reasoning and generation beyond simple speaker verification, we associate each pair with a natural language prompt (e.g., personality trait analysis, vocal expressiveness, or acoustic comparison) and a corresponding target description. Prompts are sampled from a curated template set, while reference texts are derived from manual annotations, model-generated responses, or crowd-sourced comparisons depending on availability.

Target descriptions are generated automatically using Llama-8B. We found Llama-8B strikes a practical balance between performance and speed, producing more accurate and interpretable outputs than Llama-1B and Llama-3B, while being significantly faster and more efficient to run than 70B-Instruct [20].

Unlike prior work that relies on downstream ASR or captioning labels, our dataset emphasizes open-ended speaker comparison tasks. This allows our model to generalize to more abstract or qualitative attributes such as vocal tone, emotional nuance, and prosodic variation.

5 Experiment

5.1 Baseline Model

The speaker embedding dimension is fixed to 192. And we choose SALMONN as our baseline model. For fairness, we used the same prompts for both SALMONN and SALMONN-Lite.

SALMONN configuration as following:

Architecture: Dual encoders (Whisper-Large-v2 for speech, BEATs for audio) fused via window-level Q-Former, connected to a frozen Vicuna-13B LLM with LoRA adapters

Training Method: Cross-modal pre-training + instruction tuning + activation tuning

Parameters: 13B total (33M trainable: Q-Former + LoRA)

The inference accuracy is around 97% over 1000 test data pairs

5.2 Experimental Setup

We train our model using a frozen GPT-2 language model and a frozen ECAPA-TDNN speaker encoder. The only trainable parameters are those in the Q-Former module and the linear projection layer that aligns Q-Former outputs with the GPT-2 embedding space. The Q-Former consists of two 4-layer BERT-style transformers: one for audio feature encoding and another for prompt token processing. All models are implemented in PyTorch and initialized using HuggingFace Transformers and SpeechBrain.

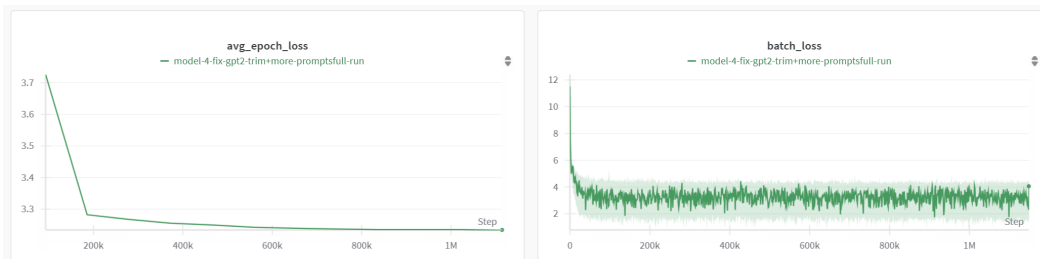
For each training sample, speaker embeddings are extracted via ECAPA, paired with a randomly selected natural language prompt, and tokenized using the GPT-2 tokenizer with added special tokens [SPK1] and [SPK2]. We use a batch size of 30 and train the model for 100 epochs using AdamW with a learning rate of $1e-4$. The learning rate follows a linear warmup schedule for the first 3 epochs, then decays following a cosine schedule. A maximum gradient norm of 1.0 is enforced for stability.

The training objective is standard next-token prediction over the target text, with a cross-entropy loss applied after concatenating the prefix embeddings from the Q-Former with the prompt token embeddings. Padding tokens are ignored in the loss. WandB is used for logging loss and prediction samples.

To encourage generalization, the dataset includes a balanced mixture of same-speaker and different-speaker pairs and leverages 10 manually designed prompts covering traits such as vocal expressiveness, acoustic variability, and personality dimensions.

5.3 Result Analysis

We assess the quality of generated speaker comparisons across both same-speaker and different-speaker cases. As illustrated in the qualitative examples, our model captures basic acoustic cues such as pitch and loudness, but often produces brief or under-specified descriptions (e.g., “Speaker 2 has more higher pitch”). Though this is a good start, we do expect more for the model. For instance the expected outputs contain detailed and structured descriptions involving phonetic features like glottalization and creakiness.



Example Comparison with ground truth (same speaker):

Generated: The voices are identical.

Expected: These two speakers are the same person. They share identical voice traits.

Prompt: Does one speaker sound more emotionally controlled or restrained than the other?

Example comparison with ground truth (difference speakers):

Prompt: Compare these two speakers and determine if they are the same person. Analyze their voice traits including pitch, timbre, resonance, and speaking patterns.

Table 1: Sample Generated Response by SALMONN-Lite

Is Same Speaker	Generated Response
No	identify the difference between the first and second speaker. The first speaker has a higher number of vocalizations than the second speaker. The second speaker has a higher number of crassness than the first speaker.
No	Speaker 1 is louder, speaker 2 has more higher pitch.
Yes	they are similar to same voice
Yes	the same traits
Yes	They are identical.

Generated: Speaker 1 is louder, speaker 2 has more higher pitch.

Expected: The two speakers have different voice quality characteristics, which can be described as follows: (1) The first speaker has a high degree of glottalization. This means that their vocal folds are engaged in a manner that produces a creaky or breathy sound. (2) The second speaker has a moderate degree of creakiness. This means that their vocal folds are engaged in a way that produces a creaky sound.

Analysis

Compared to SALMONN, which achieves 97% accuracy, our model reaches 60% accuracy on the VoxCeleb evaluation set. However, it is worth to note that SALMONN performs binary classification (yes/no) on speaker identity, while our model is tasked with open-ended natural language generation describing how two speakers differ or align across multiple acoustic and personality dimensions. This makes our task significantly harder and inherently more subjective to evaluate.

Moreover, SALMONN relies on a large instruction-tuned language model (e.g., Vicuna 13B) with LoRA adapters, trained through a multi-stage pipeline involving pretraining, instruction tuning, and activation tuning. In contrast, our system uses a unfrozen GPT-2 decoder (124M parameters) and updates only the Q-Former, projection layers, and part of GPT-2 weights. This design enables efficient training and inference with minimal hardware requirements while maintaining strong performance on complex descriptive tasks.

6 Limitations

While our approach is lightweight and efficient, it presents several limitations. First, the model uses fixed-size speaker embeddings without access to raw temporal audio features, limiting its ability to capture fine-grained prosodic or contextual dynamics such as hesitation, emphasis, or rhythm. Second, the backbone language model is GPT-2, which is significantly smaller than modern instruction-tuned LLMs and may restrict the fluency and expressiveness of the generated descriptions. Finally, our reliance on a fixed set of handcrafted prompts may introduce bias and limit generalization to unseen task instructions.

7 Future Directions

1. **Fine-Tune Casual Lanaguage Model:** The audio description between speaker pairs is a key advantage of our model, and the quality of the speakers’ description directly affects the performance. A specialized Casual Lanaguage Model tuned by specified list of prompt instruction set with tailored responses could boost a lot of the model’s performance. However, this means a lot of extra data-gathering and pre-processing. Our future work would involve fine-tuning our GPT2 configuration, given that it is frozen right now.

2. **Try constrative-based loss functions:** A contrastive loss could be applied to ensure that generated descriptions are more similar for same-speaker pairs and more distinct for different-speaker pairs, encouraging the Q-Former to encode more discriminative and meaningful speaker features. This could be implemented using cosine similarity between text embeddings of the generated and reference descriptions, or by incorporating contrastive supervision across batches.

3. **Include different tasks to SALMONN-Lite:** Based on our SALMONN-Lite framework, the model is very easily to add on additional task while doing speaker verification to generate more descriptive information on the speakers. Our future direction includes adding different personality assessments such as MBTA and HEXACO along with the training data to help model understand speakers other personalities.

8 Team Work Breakdown

Lainey Liu: Salmonn model inference, Q-former implementation, Helped with model training, Report

Bowen Ding: Data generation, Model Implementation, Model Training, Model Inference, Report

Aria Wu: Helped with Salmonn model inference and model training, Ollama and LM-BFF exploration

References

- [1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.
- [2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
- [3] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. IEEE ICASSP*, 2014, pp. 4052–4056.
- [4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE ICASSP*, 2018, pp. 5329–5333.
- [5] J. S. Chung *et al.*, "In defence of metric learning for speaker recognition," *arXiv preprint arXiv:2003.11982*, 2020.
- [6] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," *arXiv preprint arXiv:2005.07143*, 2020.
- [7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.
- [8] H. Zeinali, S. Wang, A. Silnova, P. Matejka, and O. Plchot, "BUT system description to VoxCeleb speaker recognition challenge 2019," *arXiv preprint arXiv:1910.12592*, 2019.
- [9] J.-W. Jung, H.-S. Heo, J.-H. Kim, H.-J. Shim, and H.-J. Yu, "RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification," in *Proc. Interspeech*, 2019, pp. 1268–1272.
- [10] T. Zhou, Y. Zhao, and J. Wu, "ResNeXt and Res2Net structures for speaker verification," in *2021 IEEE SLT Workshop*, pp. 301–307.
- [11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE CVPR*, 2018, pp. 7132–7141.
- [12] S.-H. Gao *et al.*, "Res2Net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2021.
- [13] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in TDNNs and frequency positional information in 2D ResNets to enhance speaker verification," *arXiv preprint arXiv:2104.02370*, 2021.

- [14] T. Liu, R. K. Das, K. A. Lee, and H. Li, “MFA: TDNN with multi-scale frequency-channel attention for text-independent speaker verification with short utterances,” *arXiv preprint arXiv:2202.01624*, 2022.
- [15] A. Vaswani *et al.*, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
- [16] Z. Dai *et al.*, “Transformer-XL: Attentive language models beyond a fixed-length context,” *arXiv preprint arXiv:1901.02860*, 2019.
- [17] Z. Yang *et al.*, “XLNet: Generalized autoregressive pretraining for language understanding,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [18] H. Zhou *et al.*, “Informer: Beyond efficient transformer for long sequence time-series forecasting,” in *Proc. AAAI*, 2021.
- [19] A. Baali *et al.*, “Phoneme-Debiasing Attention Framework (PDAF) for Speaker Verification,” 2024.
- [20] H. Tang *et al.*, “SALMONN: A Multimodal Large Language Model with Auditory Capabilities,” 2024.
- [21] A. Radford *et al.*, “Robust speech recognition via large-scale weak supervision,” OpenAI, 2022. [Online]. Available: <https://cdn.openai.com/papers/whisper.pdf>
- [22] Y. Zhang *et al.*, “MFA-Conformer: Multi-Scale Feature Aggregation Conformer for Automatic Speaker Verification,” in *Proc. Interspeech*, 2022.
- [23] B. Zhang, H. Cui, V. Nguyen, and M. Whitty, “Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead,” *Sensors (Basel)*, vol. 25, no. 7, p. 1989, 2025. doi:10.3390/s25071989.