
AUDIO EDITING FEATURES AS USER-CENTRIC PRIVACY DEFENSES AGAINST LARGE LANGUAGE MODEL (LLM)-BASED EMOTION INFERENCE ATTACKS

Mohd. Farhan Israk Soumik

School of Computing
Southern Illinois University
Carbondale, IL, 62901
mohdfarhanisrak.soumik@siu.edu

W.K.M Mithsara

School of Computing
Southern Illinois University
Carbondale, IL, 62901
malithi.mithsara@siu.edu

Abdur R. Shahid

School of Computing
Southern Illinois University
Carbondale, IL, 62901
shahid@cs.siu.edu

Ahmed Imteaj

School of Computing
Southern Illinois University
Carbondale, IL, 62901
imteaj@cs.siu.edu

February 11, 2025

ABSTRACT

The rapid proliferation of speech-enabled technologies, including virtual assistants, video conferencing platforms, and wearable devices, has raised significant privacy concerns, particularly regarding the inference of sensitive emotional information from audio data. Existing privacy-preserving methods often compromise usability and security, limiting their adoption in practical scenarios. This paper introduces a novel, user-centric approach that leverages familiar audio editing techniques, specifically pitch and tempo manipulation, to protect emotional privacy without sacrificing usability. By analyzing popular audio editing applications on Android and iOS platforms, we identified these features as both widely available and usable. We rigorously evaluated their effectiveness against a threat model, considering adversarial attacks from diverse sources, including Deep Neural Networks (DNNs), Large Language Models (LLMs), and reversibility testing. Our experiments, conducted on three distinct datasets, demonstrate that pitch and tempo manipulation effectively obfuscates emotional data. Additionally, we explore the design principles for lightweight, on-device implementation to ensure broad applicability across various devices and platforms.

1 Introduction

The interconnection between human emotion and audio is profound and intricate. Audio can evoke, amplify, and manipulate feelings through tones, rhythms, and frequencies [1]. Acoustic details and voice inflections convey emotions more effectively than words. The prevalence of audio in our daily lives has increased due to advancements in modern communication technologies, particularly high-speed internet and mobile networking like 4G and 5G. These technologies have enabled the transmission of superior audio quality and seamless transmission, which has become the cornerstone of platforms like video conferencing, podcasts, and Voice-over-IP(VoIP) services. Video conferencing and streaming platforms like Zoom, Google Meet, Twitch, and social media have contributed to this expansion. AI-powered Personalized Virtual Assistants(e.g Alexa, Siri) and Large Language Models (LLMs) are gaining ground, making voice commands and interactions more ubiquitous while increasing the significance of audio as a data format[2]. We are slowly embracing the world of virtual reality(VR), augmented reality(AR), and immersive technologies, which are considered Extended Reality [3][4]. Audio is a critical component for creating fully immersive experiences. Spatial audio, which lets users perceive sound directionally in 3D, has become crucial in AR and VR environments for more

realistic and engaging interactions. In VR gaming, dynamic audio effects synchronize sound with virtual objects and movements, enhancing immersion. Machine learning optimizes audio processing for immersive technologies, including real-time noise cancellation, voice recognition, and personalized soundscapes.

As new disruptive technologies proliferate, audio gains relevance but also brings emerging security concerns. Audio recordings reveal content and sensitive private information, like emotions or geographic locations, from background noise and vocal cues, emphasizing audio privacy [5]. Systems like Siri and Alexa use Speech Emotion Recognition (SER) to infer users' affective states for applications like mental health monitoring and interactive entertainment[6]. However, these systems are vulnerable to inference attacks that extract sensitive attributes like identity and gender from audio signals[6]. Even Federated Learning, considered more secure, has been shown to be susceptible to inference attacks on audio data[7]. Video conferencing platforms are also susceptible to attacks that extract personal attributes from audio, video, drawing, and writing data[8, 9]. Emerging VR devices present additional challenges due to their extensive sensor integrations, including audio. Attackers can exploit vocalizations in VR environments to fingerprint users and infer sensitive attributes like age, emotion, and personality using machine learning[10, 11, 12].

To mitigate the inference of sensitive attributes and information from speech data, researchers have proposed diverse privacy-preserving methods. These include sanitizing sensitive features to obscure private information, employing cryptographic techniques, and incorporating privacy-by-design principles to balance privacy and utility. Techniques such as differential privacy, disentangled representation learning, and deep neural network-based obfuscation have been explored to protect against attribute inference attacks, including gender and emotion recognition [13, 14, 15]. Frameworks leveraging advanced methods like hierarchical noise addition, task-relevant data transformation, and genetic programming have also been proposed to ensure privacy in speech-based systems, including federated learning, voice user interfaces, and smart speakers [16]. If we closely analyze these works, it becomes evident that they primarily achieve privacy by identifying and separating sensitive features of speech audio data using machine learning models. These models often involve computationally intensive architectures, such as autoencoders, to extract and isolate sensitive attributes. Once identified, these attributes are privatized using security measures like encryption or differential privacy. While some approaches allow users to define varying levels of privacy (e.g., high, medium, low) for different types of sensitive attributes, this often requires an understanding of the technical implications of these settings. Such reliance on user expertise or awareness may limit the accessibility and usability of these solutions, especially for non-technical users.

This paper proposes a novel approach to privacy preservation in audio-based emotion recognition attacks. Unlike existing works, it focuses on usability and aims to develop a method that can be adopted by users with minimal training or technical expertise. WE envisioned that the approach will be system-wide and not confined to specific applications or platforms, such as Apple Intelligence[17]. Equally important, utilizing such features to modify audio files will have minimal impact on user acceptability, as users are already employing these features for fun and utility, thereby blurring the gap between usability and privacy protection.

To achieve this goal, we begin our study by analyzing audio editing applications available on the Google Play Store[18] and Apple App Store[19] to identify common audio editing features. The rationale for this step lies in the fact that billions of people already use devices integrated into either the Android[20] or iOS[21] ecosystems. We hypothesize that leveraging familiar tools and workflows will enhance the usability and adoption of our proposed approach. Our analysis reveals that many of these applications include features to modify pitch and tempo values through intuitive means, such as sliders. Building on this insight, we explore the potential of using systematic pitch and tempo manipulation as the foundation of a user-centric privacy-preservation mechanism. To rigorously evaluate the feasibility of this approach, we design a threat model that incorporates a diverse range of attack vectors, including various AI models based on Deep Neural Networks (DNNs), Large Language Models (LLMs), and reversibility testing using audio editing applications. In the evaluation, we utilize three distinct datasets to assess the generalizability of the proposed features in protecting privacy. Furthermore, we investigate potential biases (e.g., demographic biases) in the effectiveness of privacy protection and evaluate the approach's resilience against various attack types. Our extensive experiments demonstrate promising results, indicating that pitch and tempo manipulation can serve as an effective, user-friendly privacy-preserving mechanism against emotion recognition attacks. These findings pave the way for the development of practical, usable privacy-preserving defenses to safeguard sensitive data from audio files. In a nutshell, the main contributions of our paper are as follows.

User-Centric Privacy Against Emotion Inference Attacks. This paper introduces a novel, user-centric approach to privacy preservation against audio-based emotion recognition attacks with focus on user's usability through existing tools and concepts.

Leveraging Familiar Audio Features as Basis for Privacy-Preservation. The proposed approach leverages intuitive audio-editing features such as pitch and tempo manipulation, widely available in existing ecosystems.

Extensive Threat Model for Emotion Inference Attacks. A rigorous threat model is designed, encompassing diverse attack vectors, including Deep Neural Networks (DNNs), Large Language Models (LLMs), and reversibility testing, to rigorously evaluate the proposed approach.

Experimental Validation and Generalizability. Extensive experiment conducted by carrying out pitch and tempo manipulation directly via a popular app on three different datasets demonstrating the effectiveness, generalizability, and robustness of the proposed method.

2 Related Work

Emotion Inference Attacks on Audio Data: Emotion inference attacks significantly threaten the privacy and security of systems that process emotional data. These attacks exploit machine learning models to infer sensitive, emotional states or related attributes without user consent, raising concerns in applications like virtual assistants, healthcare, and surveillance. Understanding their vulnerabilities to such attacks is critical as emotion recognition technologies become increasingly integrated into everyday systems. Most previous studies on attribute inference attacks in speech emotion recognition focus on scenarios where adversaries exploit model outputs to deduce private speaker attributes, such as identity, gender, or demographic traits, beyond the intended emotion classification[13, 22, 23]. Our study focuses on changing the tempo and pitch in audio by using up-tempo based on usable privacy and eventually changing the emotion of the video. Judith et al. [24] propose a method that involves altering pitch in simple tones across different frequency ranges to measure people's emotional states by recording their body movements. Alica et al. [25] investigate the impact of tempo (measured in beats per minute) and rhythmic units (patterns like whole, half, eighth, and sixteenth notes) on listeners' emotional perceptions and mood regulation. It aims to understand how these factors induce positive or negative emotional states. Zhiyuan et al[26] introduces a novel class of adversarial attacks on voice-controlled systems (VCS), named SMACK, which manipulates semantic speech attributes (like prosody) to evade state-of-the-art defenses by combining genetic algorithms with gradient estimation for prosody optimization.

Emotion Recognition Through Conventional Machine Learning Approaches and Large Language Models on Audio Data: Mohd et al.[27] explore emotion recognition through audio and textual modalities using a Bi-LSTM model for audio data and a fine-tuned DistilRoBERTa model for textual data and they utilize the SAVEE, RAVDESS, and TESS datasets for analysis. Theresa et al. [28] propose a novel approach, MATS2L, to improve emotion recognition using physiological signals through self-supervised learning (SSL) and 1DCNN and using the WESAD dataset. Mandeep et al.[29, 30] explore emotion recognition by processing audio spectrograms and video frames using deep learning architectures, including a 3D CNN for video frames and a CNN+RNN for audio spectrograms. Their work utilizes the IEMOCAP dataset, which comprises 12 hours of audiovisual data from 10 actors expressing various emotions, such as anger, happiness, sadness, and neutrality. When considering conventional methods, machine learning techniques have reached a significant milestone in accurately identifying emotions, demonstrating remarkable advancements in the field[29, 30]. However, regarding utilizing large language models (LLMs) for emotion recognition, only a few studies have been conducted, leaving this area relatively under-explored [31, 32]. Zehui et al.[31] explore a novel approach to integrating speech characteristics into Large Language Models (LLMs) for emotion recognition. The study focuses on enhancing the multi-modal capabilities of LLMs without requiring architectural modifications by converting audio features such as volume, pitch, and speaking rate into natural language descriptions. The researchers evaluate various LLMs (e.g., LLaMA-2, LLaMA-3, Phi-3) using LoRA fine-tuning and IEMOCAP and MELD datasets. Yaoxun et al. [32] introduce a novel framework, SECap, for describing speech emotions using natural language, moving beyond traditional classification into predefined emotion categories. They utilize HuBERT as the audio encoder to extract robust speech features, while the Bridge-Net is employed to process these features, separating emotion-related acoustic information from content. LLaMA is then used to generate coherent and human-like emotion captions. The study utilizes the EMOSpeech dataset, which consists of 41.6 hours of annotated speech data, including corresponding emotion captions, transcriptions, and labels.

Privacy-Preserving Mechanisms to Defend Against Emotion Inference Attacks: Privacy-preserving mechanisms aim to mitigate these threats by sanitizing sensitive features, applying cryptographic techniques, or incorporating privacy-by-design principles. These methods ensure that emotion recognition systems retain functionality while protecting users' emotional data from unauthorized inference, balancing utility and privacy in sensitive applications. Haijiao et al.[13] address privacy challenges in federated learning (FL) for Speech Emotion Recognition by introducing a Gradient-level Hierarchical Differential Privacy (GHDP) framework by employing gradient normalization, clipping significant gradients, and adding hierarchical noise to early model layers during backpropagation to mitigate attribute inference attacks, such as gender prediction. Ranya et al.[14] explore privacy challenges in voice user interfaces (VUIs), which are widely used in smartphones, home assistants, and IoT devices. The focus is on defending against attribute inference attacks that exploit deep acoustic models to infer private attributes like gender or emotion. The authors propose a novel framework utilizing disentangled representation learning to filter sensitive information while retaining

utility for primary tasks such as speech recognition. Wang et al.[15] proposes a novel privacy-preserving framework for audio-based applications by implementing a deep neural network (DNN)-based obfuscator to transform raw audio into a privacy-preserving format, retaining only task-relevant data without taking sensitive attributes like gender, emotion, or identity. This allows users to authorize only specific tasks while obfuscating unauthorized information. Brian et al[16] introduce DARE-GP (Defeating Acoustic Recognition of Emotion via Genetic Programming), a novel framework to protect user privacy from unauthorized speech emotion recognition (SER) systems, such as those embedded in smart speaker voice assistants (e.g., Amazon Echo, Google Home).

Observations: Current methods for defending against emotion inference attacks are heavily reliant on computationally intensive models, such as deep learning and disentangled representation learning, which focus on separating sensitive features like emotions. While effective, these approaches often overlook usability and adaptability, and hence limits their practical deployment for non-technical users. Furthermore, techniques like pitch and tempo manipulation, which are intuitive and accessible, remain underexplored for privacy preservation. Current frameworks also lack emphasis on cross-platform applicability and usability, particularly in diverse systems like virtual assistants and VR devices. Our work addresses these limitations by leveraging familiar audio-editing features to design lightweight, user-centric privacy mechanisms that balance effectiveness and accessibility.

3 Threat Model

Assumption on Target Systems: We consider target systems that interact with speech data in various capacities. These include any system that records, stores, processes, or provides third-party access to speech data. Examples of such systems include virtual assistants, such as Alexa[33], Siri[34], and Google Assistant[35], which process user input to perform tasks or provide services, and large language model-based chatbots, such as ChatGPT[36] and Gemini[37], which may handle audio data as part of their multimodal interaction capabilities. Video conferencing platforms, such as Google Meet[38] and Zoom[39], that record or store audio for communication purposes are also included. Also, cloud-based audio applications, which store user data in cloud infrastructures, pose risks such as unauthorized access, data breaches, and malicious exploitation. Emerging technologies like Meta Smart Glasses[40], which integrate audio for real-time communication, transcription, or augmented reality, process and potentially store speech data, raising privacy concerns due to their pervasive nature and continuous data collection.

Adversaries and Attack Vectors: The mentioned systems can be targeted by various adversaries with different motivations and methods. External hackers can breach cloud storage systems or intercept communication channels to extract speech data for unauthorized analysis or dissemination. Application developers also pose a potential risk when they have access to users' audio data. Developers may misuse this data for purposes beyond original intent for monetization through emotion recognition analysis or sharing it with third parties without the explicit consent of users. Similarly, insider threats, such as colluding employees or individuals within an organization, with access to the sensitive data, may misuse their access for personal or financial gain or collaborate with external parties to exploit the data.

Assumption on Attacker's Knowledge: We assume that the attackers possess significant knowledge and resources to compromise the privacy of speech data. This includes, knowledge of the application used to manipulate the speech data, including its features or method used; expertise and resources to develop or access advanced AI models specifically designed for inference attacks; and access to state-of-the-art LLMs like GPT-4o, which can analyze audio files to infer sensitive data.

4 The Proposed Framework

4.1 Design Goals

The core challenge in preserving privacy against the threat model lies in designing a mechanism that effectively protects sensitive emotional information while maintaining usability, adaptability, and compatibility across diverse systems and devices. This necessitates addressing several critical factors, including computational efficiency, user accessibility, and system-wide applicability. These goals include: (*Lightweight On-Device Implementation*) The

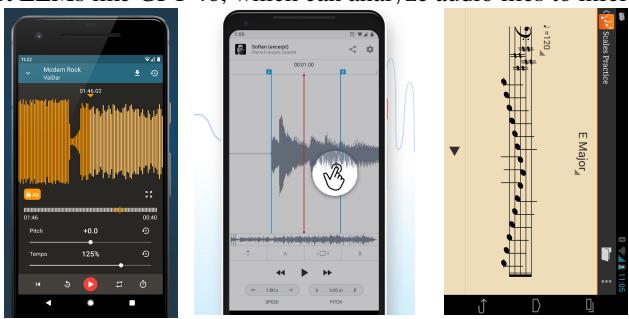


Figure 1: Interface of audio editing apps: Up Tempo, AudioStretch, and Scales Practice. (Images taken from the apps' sites.)

privacy-preserving mechanism should be computationally lightweight so that it can be implemented directly on user devices without requiring extensive processing power or memory. This is critical for devices with limited resources, such as smartphones, smart glasses, or IoT devices. (*Usability*) The solution should prioritize ease of use which will

Table 1: Example popular audio editing apps and their key features, including pitch and tempo manipulation.

App Name	Platform	Key Features	Popularity
Up Tempo	Android/iOS	Pitch adjustment, tempo control	1 Million+ downloads on Android, Rating 5 on iOS
Scales Practice	Android/iOS	Pitch adjustment, tempo control	100K+ Downloads on Android, 4.8 rating on iOS
AudioStretch	Android	Pitch adjustment, speed control	1 Million+ downloads with 3.5 average rating at playstore.
BandLab	Android	Pitch and Tempo shifting	50 Million+ downloads with 4.7 rating at playstore
Garage Band	iOS	Tempo and Pitch Adjustments	98k+ reviews with 4.1 rating at Appstore

Data is based on Google Play Store and Apple App Store ratings and download counts as of Nov. 2024.

allow users to adopt privacy-preserving features with minimal technical expertise or training. This involves intuitive interfaces and seamless integration with existing workflows to reduce barriers to adoption. (*Adaptability*) It is highly desirable that the mechanism should be adaptable to a wide range of applications and scenarios, including virtual assistants, video conferencing platforms, and cloud-based services. Furthermore, it should accommodate varying data distribution and device capabilities for consistent performance across heterogeneous systems. Furthermore, it should be compatible with a wide range of devices and platforms, including Android, Windows, iOS, etc. (*Robustness Against Attacks*) The design must be resilient to various attack vectors, including reversibility attempts, AI-based inference models, and adversaries leveraging large language models (LLMs).

4.2 The Usable Privacy-Preserving Mechanism

Our methodology to achieve the above mentioned design goals is rooted in usable security and privacy. Our objective is to develop privacy-preserving defense mechanisms based on features that users are already familiar with, which will enable intuitive adoption and minimizes the learning curve. In our first step of the research, we investigate popular audio editing features that are widely available and easily accessible to users. To accomplish this, we analyzed audio editing applications available on both the Google Play Store and Apple App Store. We found that pitch and tempo adjustments are among the most commonly available and widely used features in popular audio editing applications. Using this observation, we next explore whether the adjustment of pitch and tempo provides a privacy protection against sensitive information inference attacks from audio. We hypothesize that by integrating their manipulation into the design of privacy preserving mechanism, we can achieve the dual goals of usability and privacy preservation. Since users are already familiar with these features, they are more likely to adopt and use them effectively as a privacy-preserving mechanism. To determine whether pitch and tempo manipulation can genuinely provide privacy, we designed and conducted a series of experiments.

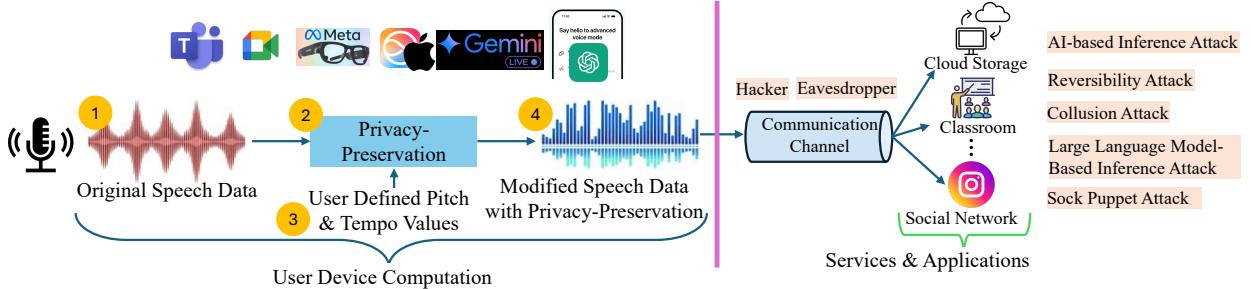


Figure 2: The Privacy-Preserving mechanism integrated system model.

Our envisioned privacy-preserving system model is depicted in Fig 2, illustrating the flow from original speech data to modified data and the associated risks. Initially, the original speech data such as that from video conferencing platforms like Microsoft Teams and Google Meet or recordings captured through devices like Meta Ray-Ban, Apple Siri, or Gemini/ChatGPT undergoes user defined modifications. These modifications involve altering pitch and tempo values to different levels using practical applications as shown in Fig 1. Changing pitch and tempo in speech directly affects the emotion conveyed in the audio file. For example, if an audio file contains a recording in a friendly or happy tone, altering the pitch and tempo can convert it to another emotion, such as anger or disgust. The modified speech data is then transmitted through various communication channels, including user devices, computational platforms, cloud storage, and services or applications such as classrooms and social networks. Throughout this process, the framework addresses several potential privacy threats, including interception by hackers or eavesdroppers, AI-based inference attacks to extract sensitive information, reversibility attacks that attempt to reconstruct the original data, collusion attacks involving multiple entities, and inference attacks leveraging large language models. Additionally, there is a risk of sock puppet attacks, where malicious entities use fake identities to exploit the system.

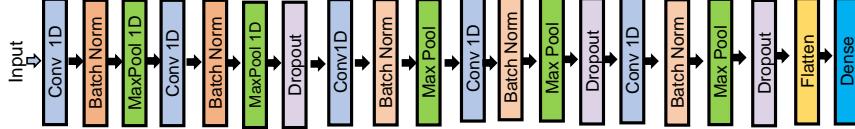


Figure 3: 1D-CNN Model Architecture

4.3 Trained Threat Models

As our potential attacker possesses the capability of using deep learning model including immensely powerful LLM models for the aforementioned inference attack, for attack simulation first we designed a 1-D CNN model from scratch and trained it using the RAVDESS, CREMA-D and three prominent dataset for emotion recognition. The reason for using 1-D CNN is that not only it is lightweight which increases the research scope in that direction but also in recent researches with audio and speech data it showed impressive performance as demonstrated in [14]. As expected, we achieved a significantly high 97% of test accuracy for the model. Fig 3 shows the architecture of our 1 dimensional convolutional neural network model.

5 Experiment and Results

Setup and Datasets: In this study, we analyzed popular audio editing applications, such as Uptempo, available on the Google Play Store and Apple App Store, to identify the common audio editing features offered by these platforms. Our objective is to assess the potential of these features as privacy-preserving tools, specifically in the context of protecting user privacy against speech emotion recognition systems. Within our threat model, we explore two primary approaches for executing attacks to evaluate the effectiveness of these features. The first method involves randomly altering pitch and tempo values while selecting male and female actors from three datasets to examine the impact on privacy. The second method involves systematically adjusting pitch and tempo across six distinct levels to create a range of variations. For pitch, the selected values are 0.0, +4, -4, +8, and -8, while tempo is varied at 60, 80, 100, 120, and 140. These combinations yield a total of 25 unique attack scenarios. For instance, when the pitch is set to 0.0, it is paired sequentially with tempo values of 60, 80, 100, 120, and 140. Similarly, for a pitch value of -4, the same tempo values are applied. This systematic approach is consistently implemented across all pitch variations to thoroughly examine their combined effects. We use three emotion speech datasets, specifically the RAVDESS [41], CREMA-D[42], and TESS[43] datasets. The details of these datasets are presented in Table 2.

Table 2: Summary of Emotion Speech Datasets

Dataset	Total Files	Actors	Age Range	Emotions	File Description
RAVDESS[41]	1440	24 (12 male, 12 female)	-	Calm, Happy, Sad, Angry, Fearful, Surprised, Disgust	60 trials × 24 actors; 16-bit, 48 kHz audio
CREMA-D[42]	7442	91 (48 male, 43 female)	20-74	Anger, Disgust, Fear, Happy, Neutral, Sad (Low intensity)	12 sentences × 91 actors; multi-modal data
TESS[43]	2800	2 (female)	26, 64	Anger, Disgust, Fear, Happiness, Pleasant Surprise, Sadness, Neutral	200 words × 2 actresses; carrier phrase: 'Say the word...'

5.1 Randomized Approach

To evaluate our proposed attack and defense model using the Up Tempo app, we randomly altered the pitch and tempo for different categories of emotion of the same user with different combinations that are fed to the attacker's models as inputs. To examine gender-related biases, we selected one male and one female actor from each of the aforementioned datasets and analyzed their pitch variations. Table 3 shows the result for Actor 23 of REVDESS dataset on randomly selected tempo and pitch values. The results clearly demonstrate that changed pitch and tempo is creating significant effects on attacker's inference model's performance in every case from 1-D CNNs to highly capable LLM model such as GPT-4o. Similarly, Table 4 shows the outcome for actor 16 and we observe that just like the previous actors with the randomized change in tempo and pitch values, attacker both with their 1-D CNN and GPT-4o LLM model is failing to properly extract the real emotion from the data.

Table 5 illustrates the impact of pitch and tempo variations on a male actor (Actor ID: 1001). As observed in the previous cases, the attacker consistently infers incorrect emotional information from the speech files of this actor. Following the approach taken with the RAVDESS dataset, where we transitioned from testing a male actor (Actor 23) to a female actor (Actor 16), we conducted similar tests on a female actor from the CREMA-D dataset. Table 7

highlights the effects of various random combinations of pitch and tempo on the female actor (Actor ID: 1002), with the significant impact of these variations clearly evident in the results.

Beyond the evident failure of the proposed attacker models to correctly infer emotions, a deeper analysis of the randomized approach reveals an important insight: female actors, even when drawn from different datasets and representing entirely different individuals, exhibit notable commonalities in how their emotions are affected by pitch and tempo variations. For instance, when the original emotion is anger, a combination of high pitch and high tempo often leads to a predicted emotion of fear. Similarly, for disgust, the same combination frequently results in a shift to fear. Observations from Table 3 and Table 5 further corroborate this trend. For example, increasing the tempo while keeping the pitch constant alters the predicted emotion from neutral to disgust. These findings suggest a strong correlation between pitch, tempo, and emotional expression in speech data, prompting a more structured and detailed analysis, which is presented in the next section.

Table 3: Effect of different pitch and tempo change on Male (ID 23 RAVDESS)

Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion	Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion
-9.4	72.0	neutral	neutral	Neutral	0.0	129.0	happy	neutral	Disgust
-8.4	76.0	happy	happy	Fear	0.0	130.0	fear	sad	Surprise
-8.4	100.0	neutral	happy	Neutral	0.0	138.0	angry	angry	Fear
-8.4	136.0	surprise	surprise	Neutral	0.0	142.0	disgust	disgust	Neutral
-8.3	100.0	disgust	angry	Disgust	6.3	100.0	fear	fear	Disgust
-7.8	76.0	neutral	neutral	Disgust	7.4	100.0	surprise	fear	Fear
-7.6	73.0	fear	fear	Surprise	7.8	126.0	fear	fear	Surprise
-7.5	100.0	surprise	surprise	Fear	7.9	100.0	surprise	fear	Surprise
-7.4	100.0	surprise	fear	Surprise	8.1	61.0	disgust	angry	Neutral
0.0	61.0	disgust	sad	Neutral	8.1	126.0	angry	sad	Disgust
0.0	68.0	sad	disgust	Disgust	8.7	128.0	fear	fear	Fear
0.0	70.0	sad	sad	Fear	10.6	133.0	surprise	fear	Neutral
0.0	72.0	happy	neutral	Surprise	10.9	100.0	neutral	neutral	Neutral

Table 4: Effects of pitch and tempo change on Female (ID-16 RAVDESS)

Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion	Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion
0.0	143	angry	angry	Surprise	0.0	66	neutral	happy	Disgust
0.0	63	angry	Disgust	Surprise	0.0	66	neutral	neutral	Disgust
9.7	141	fear	fear	Surprise	0.0	142	surprise	fear	Disgust
-8.9	62	sad	disgust	Surprise	7.6	133	fear	fear	Disgust
-8.1	100	surprise	fear	Surprise	-7.5	70	fear	fear	Disgust
10.3	100	fear	fear	Surprise	-5.7	100	disgust	angry	Disgust
0.0	66	fear	fear	Sad	8.9	100	angry	angry	Disgust
8.0	100	sad	sad	Sad	9.0	133	fear	fear	Anger
-7.9	100	disgust	disgust	Sad	-8.5	70	disgust	sad	Anger
8.3	130	angry	angry	Sad	0.0	132	angry	angry	Anger
-9.1	68	fear	surprise	Sad	0.0	62	fear	fear	Anger
0.0	127	angry	angry	Sad	7.6	100	sad	disgust	Anger
7.2	100	sad	disgust	Neutral	-7.0	100	angry	sad	Anger

5.2 Structured Approach

While the goal of randomized approach was to see whether variations in pitch and combo can make the attacker extract incorrect emotions, the goals of this structured section is to provide a brief analysis to find out how this emotion change happens. Unlike the randomized approach, tempo and pitch values are not selected randomly this, but with regular interval. As previously mentioned, we have used pitch value of -8,-4,0,4,8 and tempo values of 60,80,100,120,140. For each of the emotion of category we made possible combinations with the designated pitch and tempo values and figured out the attacker model's response. Here, we will use previously discussed female actor (ID-16 RAVDESS) speech audio with Neutral emotion as a test case. Table 8 demonstrates the the pitch and tempo variations on actor 16's speech with neutral emotion. And it represent how different combinations of pitch and tempo influence emotion classification by two models: a 1-D CNN (Convolutional Neural Network) and GPT-4o, a variant of GPT-4 optimized for this task. Both models occasionally agree on the emotional label (e.g., at 60 BPM and -8 pitch, both predict "sad"). Lower tempos (e.g., 60 BPM) and negative pitch shifts (-8, -4) often correlate with "sad" or "disgust." Higher tempos

Table 5: Impact of Pitch and Tempo Variations on Male Actor (ID:1001 from CREMA-D)

Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion	Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion
0.0	64	happy	neutral	Fear	0.0	65	disgust	sad	Disgust
0.0	133	disgust	sad	Fear	0.0	132	disgust	angry	Disgust
-8.5	68	disgust	disgust	Fear	-8.8	68	neutral	happy	Disgust
8.2	133	disgust	disgust	Fear	8.7	134	happy	happy	Disgust
-8.5	100	fear	fear	Fear	-7.6	100	happy	neutral	Disgust
8.3	100	fear	sad	Fear	7.7	100	disgust	angry	Disgust
0.0	64	happy	neutral	Sad	0.0	67	fear	fear	Happy
0.0	136	disgust	disgust	Sad	0.0	131	disgust	angry	Happy
-7.8	70	happy	neutral	Sad	-8.2	70	happy	neutral	Happy
8.5	132	fear	sad	Sad	8.3	131	fear	fear	Happy
-8.2	100	fear	fear	Sad	-8.6	100	happy	happy	Happy
8.1	100	disgust	disgust	Sad	7.8	100	disgust	angry	Happy
0.0	68	angry	angry	Anger	0.0	67	fear	fear	Neutral
0.0	136	fear	Anger	Anger	0.0	131	disgust	angry	Neutral
-9.1	66	angry	disgust	Anger	-8.2	70	happy	neutral	Neutral
7.4	136	fear	fear	Anger	8.3	131	fear	fear	Neutral
-7.6	100	neutral	happy	Anger	-8.6	100	happy	happy	Neutral
7.8	100	angry	angry	Anger	7.8	100	disgust	angry	Neutral

Table 6: Impact of Pitch and Tempo Variations on Female (ID-1002 CREMA-D).

Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion	Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion
0.0	63	angry	angry	Disgust	0.0	66	disgust	Neutral	Neutral
0.0	132	angry	angry	Disgust	0.0	133	happy	Neutral	Neutral
-9.1	68	angry	sad	Disgust	-8.5	68	angry	disgust	Neutral
8.8	130	fear	fear	Disgust	8.4	131	happy	happy	Neutral
-7.0	100	angry	angry	Disgust	-8.2	100	neutral	happy	Neutral
7.9	100	fear	fear	Disgust	8.5	100	happy	happy	Neutral
0.0	68	sad	sad	Fear	0.0	137	fear	disgust	Sad
0.0	131	fear	fear	Fear	-8.7	70	disgust	angry	Sad
-8.1	70	disgust	angry	Fear	9.0	134	surprise	disgust	Sad
7.0	132	angry	disgust	Fear	-7.7	100	neutral	happy	Sad
-8.2	100	neutral	sad	Fear	8.9	100	sad	sad	Sad
8.0	100	sad	disgust	Fear	0.0	68	disgust	angry	Sad
0.0	66	fear	fear	Happy	0.0	63	angry	disgust	Anger
0.0	132	happy	happy	Happy	0.0	132	angry	angry	Anger
-8.0	64	fear	fear	Happy	-9.1	68	angry	disgust	Anger
8.6	131	happy	neutral	Happy	8.8	130	fear	surprise	Anger
-7.8	100	angry	disgust	Happy	-7.0	100	angry	angry	Anger
8.5	100	happy	neutral	Happy	7.9	100	fear	fear	Anger

Table 7: Impact of Pitch and Tempo Variations on Female Actor OAF of Tess Dataset

Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion	Pitch	Tempo	1-D CNN	GPT-4o	Original Emotion
0.0	84	angry	sad	Disgust	0.0	70	disgust	Neutral	Neutral
0.0	130	angry	sad	Disgust	0.0	136	happy	happy	Neutral
-9.1	68	angry	sad	Disgust	-7.5	68	angry	angry	Neutral
8.5	133	fear	fear	Disgust	8.4	131	happy	neutral	Neutral
-6.5	100	angry	angry	Disgust	-8.2	100	neutral	happy	Neutral
7.9	100	fear	fear	Disgust	8.2	100	happy	happy	Neutral
0.0	68	sad	sad	Fear	0.0	137	fear	fear	Sad
0.0	131	sad	disgust	Fear	-8.7	70	sad	angry	Sad
-6.9	72	surprise	sad	Fear	8.00	134	disgust	neutral	Sad
7.0	132	angry	disgust	Fear	-7.7	100	neutral	happy	Sad
-8.2	100	neutral	sad	Fear	8.00	100	sad	sad	Sad
8.0	100	sad	disgust	Fear	0.0	68	disgust	angry	Sad
0.0	65	fear	fear	Happy	0.0	62	angry	disgust	Anger
0.0	132	happy	happy	Happy	0.0	132	angry	angry	Anger
-8.0	64	fear	fear	Happy	-8.3	65	angry	angry	Anger
8.6	131	happy	happy	Happy	7.5	130	fear	surprise	Anger
-7.5	123	angry	disgust	Happy	-8.0	100	angry	anger	Anger
8.5	100	happy	happy	Happy	7	100	fear	fear	Anger

(e.g., 120, 140 BPM) and positive pitch shifts (4, 8) show varied predictions like “happy,” “fear,” or “surprise.” Neutral tempo (0 BPM) shows variability in emotion predictions, with both “neutral” and other emotions like “happy” appearing.

Table 9 represents for both models predict “Sad” emotion reliably at a moderate pitch (0 or 8) and moderate tempo (60 or 80 BPM). There is significant difference between 1-D CNN and GPT-4o, especially at higher tempos (100–140 BPM), where GPT-4o often predicts more varied emotions like “happy” or “surprise.” and 1-D CNN maintains some consistency in identifying “neutral” or “sad” in high-tempo scenarios.

Table 8: Effects of Pitch and Tempo variations on **Neutral** Emotion.

Tempo	Pitch	1-D CNN	GPT-4o	Tempo	Pitch	1-D CNN	GPT-4o
60	-8	sad	sad	140	8	fear	fear
60	-4	sad	sad	140	8	fear	surprise
60	8	sad	sad	140	4	fear	fear
60	4	sad	neutral	140	0	fear	fear
60	0	sad	sad	120	-8	disgust	anger
80	-8	disgust	anger	120	-4	angry	anger
80	-4	disgust	disgust	120	-4	angry	anger
80	8	disgust	anger	120	4	angry	sad
80	4	happy	happy	120	0	angry	sad
80	0	disgust	anger	0	-8	neutral	happy
140	-8	sad	sad	0	-4	neutral	happy
140	-8	sad	neutral	0	8	neutral	neutral
140	-4	fear	surprise	0	4	neutral	neutral

Table 9: Effects of Pitch and Tempo on **Sad** Emotion

Tempo	Pitch	1-D CNN	GPT-4o	Tempo	Pitch	1-D CNN	GPT-4o
60	-8	disgust	anger	140	8	fear	fear
60	-4	sad	sad	140	4	fear	surprise
60	8	sad	sad	140	0	fear	surprise
60	4	sad	disgust	120	-8	neutral	neutral
60	0	sad	disgust	120	-4	neutral	neutral
80	-8	disgust	anger	120	8	surprise	surprise
80	-4	disgust	anger	120	4	surprise	anger
80	8	sad	sad	120	0	surprise	surprise
80	4	disgust	anger	100	-8	neutral	happy
80	0	disgust	anger	100	-4	neutral	sad
140	-8	neutral	happy	100	8	sad	sad
140	-4	happy	happy	100	4	neutral	neutral

6 Conclusion and Future Work

In this paper, we investigate the impact of modifying the pitch and tempo of speech audio data using user-accessible apps from the Google Play Store and iOS App Store. Our findings indicate that the evaluated attack models consistently failed to accurately infer emotions from the modified data, effectively providing privacy protection for sensitive user speech data. This demonstrates a practical, user-centric approach to protect privacy of sensitive data from speech audio in the face of adversaries with capabilities of state-of-the-art inference models, including multimodal large language models. The novelty of this work lies in its practical contribution to bridging the gap between usability and privacy, offering users a promising method to protect their speech data through familiar applications.

For future work, we plan to extend the scope of this research in several important directions. We plan to simulate more advanced attackers leveraging other LLMs, such as Flan-T5, Gemma-2B, and LLama, fine-tuned specifically for speech emotion recognition (SER) tasks. We also plan to integrate recent SER advancements, including universal audio representation models like Emotion2Vec2 [44] and Wave2Vec2 [45], into our attack framework, to facilitate the development of more complex and advanced attack scenarios to get better insights into the robustness and privacy implications of speech data modifications. We will also investigate the reversibility of pitch and tempo modifications using signal-processing techniques and how to mitigate that. Furthermore, we plan to assess privacy protection for additional attributes (e.g. identify or gender) beyond emotion inference.

References

- [1] Chia-Wei Li, Tzu-Han Cheng, and Chen-Gia Tsai. Music enhances activity in the hypothalamus, brainstem, and anterior cerebellum during script-driven imagery of affective scenes. *Neuropsychologia*, 133:107073, 2019.

- [2] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.
- [3] Durand R Begault and Leonard J Trejo. 3-d sound for virtual reality and multimedia. Technical report, 2000.
- [4] Fang-Yi Chao, Cagri Ozcinar, Chen Wang, Emin Zerman, Lu Zhang, Wassim Hamidouche, Olivier Deforges, and Aljosa Smolic. Audio-visual perception of omnidirectional video for virtual reality applications. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6, 2020.
- [5] Jennifer Williams, Karla Pizzi, Shuvayanti Das, and Paul-Gauthier Noé. New challenges for content privacy in speech and audio. *arXiv preprint arXiv:2301.08925*, 2023.
- [6] Basmah Alsenani, Tanaya Guha, and Alessandro Vinciarelli. Privacy risks in speech emotion recognition: A systematic study on gender inference attack. 2023.
- [7] Tiantian Feng, Hanieh Hashemi, Rajat Hebbal, Murali Annavaram, and Shrikanth S Narayanan. Attribute inference attack of speech emotion recognition in federated learning settings. *arXiv preprint arXiv:2112.13416*, 2021.
- [8] Dima Kagan, Galit Fuhrmann Alpert, and Michael Fire. Zooming into video conferencing privacy. *IEEE Transactions on Computational Social Systems*, 11(1):933–944, 2024.
- [9] Abdur R. Shahid and Ahmed Imteaj. Securing User Privacy in Cloud-Based Whiteboard Services Against Health Attribute Inference Attacks . *IEEE Transactions on Artificial Intelligence*, 5(08):3872–3885, August 2024.
- [10] Gonzalo Munilla Garrido, Vivek Nair, and Dawn Song. Sok: Data privacy in virtual reality. *Proceedings on Privacy Enhancing Technologies*, 2024(1):21–40, January 2024.
- [11] Mahdi Alkaeed, Adnan Qayyum, and Junaid Qadir. Privacy preservation in artificial intelligence and extended reality (ai-xr) metaverses: A survey, 2023.
- [12] Chamara Sandeepa, Shen Wang, and Madhusanka Liyanage. Privacy of the metaverse: Current issues, ai attacks, and possible solutions. In *2023 IEEE International Conference on Metaverse Computing, Networking and Applications (MetaCom)*, pages 234–241, 2023.
- [13] Haijiao Chen, Huan Zhao, and Zixing Zhang. Gradient-level differential privacy against attribute inference attack for speech emotion recognition. *IEEE Signal Processing Letters*, 31:3124–3128, 2024.
- [14] Ranya Aloufi, Hamed Haddadi, and David Boyle. Privacy-preserving voice analysis via disentangled representations. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, CCSW’20, page 1–14, New York, NY, USA, 2020. Association for Computing Machinery.
- [15] Wei-Cheng Wang, Sander De Coninck, Sam Leroux, and Pieter Simoens. An opt-in framework for privacy protection in audio-based applications. *IEEE Pervasive Computing*, 21(4):17–24, 2022.
- [16] Brian Testa, Yi Xiao, Harshit Sharma, Avery Gump, and Asif Salekin. Privacy against real-time speech emotion detection via acoustic adversarial evasion of machine learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 7(3), September 2023.
- [17] Apple Inc. Apple intelligence. <https://www.apple.com/apple-intelligence/>, 2024. Accessed: 2024-11-28.
- [18] Google. Google play store. <https://play.google.com/>, 2024. Accessed: 2024-11-28.
- [19] Apple Inc. Apple app store. <https://www.apple.com/app-store/>, 2024. Accessed: 2024-11-28.
- [20] Google. Android official website. <https://www.android.com/>, 2024. Accessed: 2024-11-28.
- [21] Apple Inc. ios 18 features. <https://www.apple.com/ios/ios-18/>, 2024. Accessed: 2024-11-28.
- [22] Tiantian Feng, Raghuveer Peri, and Shrikanth Narayanan. User-level differential privacy against attribute inference attack of speech emotion recognition in federated learning. *arXiv preprint arXiv:2204.02500*, 2022.
- [23] Huan Zhao, Haijiao Chen, Yufeng Xiao, and Zixing Zhang. Privacy-enhanced federated learning against attribute inference attack for speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [24] Judith Ley-Flores, Eslam Alshami, Aneesa Singh, Frédéric Bevilacqua, Nadia Bianchi-Berthouze, Ophelia Deroy, and Ana Tajadura-Jiménez. Effects of pitch and musical sounds on body-representations when moving with sound. *Scientific reports*, 12(1):2676, 2022.
- [25] Alicia Fernández-Sotos, Antonio Fernández-Caballero, and José M Latorre. Influence of tempo and rhythmic unit in musical emotion regulation. *Frontiers in computational neuroscience*, 10:80, 2016.

- [26] Zhiyuan Yu, Yuanhaur Chang, Ning Zhang, and Chaowei Xiao. {SMACK}: Semantically meaningful adversarial audio attack. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3799–3816, 2023.
- [27] Mohd Ariful Haque, Roy George, Rakib Hossain Rifat, Md Shihab Uddin, Marufa Kamal, and Kishor Datta Gupta. A cross-domain exploration of audio and textual data for multi-modal emotion detection. In *Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '24, page 375–381, New York, NY, USA, 2024. Association for Computing Machinery.
- [28] Theresa Goetz, Pulkit Arora, F. X. Erick, Nina Holzer, and Shrutiika Sawant. Self-supervised representation learning using multimodal transformer for emotion recognition. In *Proceedings of the 8th International Workshop on Sensor-Based Activity Recognition and Artificial Intelligence*, iWOAR '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [29] Gaoyun Li, Yong Liu, and Xiong Wang. Speech emotion recognition based on 1d cnn and mfcc. In *2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT)*, pages 956–960, 2023.
- [30] Mandeep Singh and Yuan Fang. Emotion recognition in audio and video using deep neural networks. *arXiv preprint arXiv:2006.08129*, 2020.
- [31] Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances. *arXiv preprint arXiv:2407.21315*, 2024.
- [32] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. Secap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19323–19331, 2024.
- [33] Amazon. Alexa official website. <https://www.alexa.com/>, 2024. Accessed: 2024-11-28.
- [34] Apple Inc. Siri by apple. <https://www.apple.com/siri/>, 2024. Accessed: 2024-11-28.
- [35] Google. Google assistant official website. <https://assistant.google.com/>, 2024. Accessed: 2024-11-28.
- [36] OpenAI. Chatgpt official website. <https://chatgpt.com/>, 2024. Accessed: 2024-11-28.
- [37] Google. Gemini by google. <https://gemini.google.com/>, 2024. Accessed: 2024-11-28.
- [38] Google. Google meet. <https://meet.google.com/>, 2024. Accessed: 2024-11-28.
- [39] Zoom Video Communications, Inc. Zoom video communications. <https://www.zoom.com/>, 2024. Accessed: 2024-11-28.
- [40] Meta Platforms, Inc. Meta smart glasses. <https://www.meta.com/smart-glasses/>, 2024. Accessed: 2024-11-28.
- [41] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.
- [42] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.
- [43] M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS), 2020.
- [44] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation, 2023.
- [45] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.