

**PREDICTIVE ANALYSIS**

Final report

**FOR**

**EMO-FAKE**

|  |  |  |
| --- | --- | --- |
| **Specialization** | **SAP ID** | **Name** |
| **AIML .H** | **500109971** | **Girish** |
| **AIML .H** | **500109497** | **Yash kesarwani** |
| **AIML .H** | **500093654** | **Aryan Singh Payaal** |

**Submitted By**

*Under the guidance of*

**School Of Computer Science**

**UNIVERSITY OF PETROLEUM & ENERGY STUDIE**

|  |  |  |
| --- | --- | --- |
| Table of Content | | Page no. |
| 1 | Abstract | 3 |
| 2 | Introduction | 4 |
| 3 | Literature review | 5 |
| 4 | Problem statement | 6 |
| 5 | Objectives | 6 |
| 6 | Methodology | 7-9 |
| 7 | PERT chart | 10 |
| 8 | Conclusion | 10 |

# ABSTRACT:

In this study, we focus on EmoFake Detection (EFD). EmoFake (EF) is a class of audio deepfake that alters the source emotion to another emotion. Firstly, we hypothesize that multilingual speech foundation models (SFMs) will be the most effective for EFD due to their knowledge of diverse pitches, tones, intensities, and so due to their pre-training in multiple languages. To validate our hypothesis, we perform a comprehensive comparative study of various state-of-the-art (SOTA) SFMs including multilingual, monolingual as well as speaker recognition. Our experiments validated our hypothesis. To our end, secondly, we experimented with fusion of SFMs. We propose a novel framework that captures the complementary strengths of SFMs for improved EFD. We show that fusion of multilingual SFMs obtain the topmost performance in comparison to invididual SFMs and baseline fusion techniques as well as show SOTA performance in comparison to previous SOTA work

# INTRODUCTION:

EmoFake (EF) represents an advanced and nuanced type of audio deepfake technology that modifies the emotional tone of a speaker’s voice while maintain- ing the speaker’s identity and linguistic content. Un- like traditional audio deepfakes, which primarily alter speaker identity or create entirely synthetic speech con- tent, EF specifically targets the emotional component of speech. This allows for imperceptible yet impactful changes to the emotional expression within audio, mak- ing the manipulated speech appear natural and realistic. Such shifts in perceived emotion, although subtle, can have substantial implications, as they may distort the intended meaning of communication, lead to misunder- standings, and impact the listener’s emotional response. The manipulation of emotional expression in speech carries notable risks, especially in fields where the au- thenticity of emotional cues is vital. In contexts such as interpersonal communication, media, and informa- tion dissemination, altering the emotional tone of audio content can have serious consequences. Misuse of EF technology could facilitate psychological manipulation, distort perceptions, and erode trust in audio-based com- munication. It can also be leveraged to spread misin- formation, influence social interactions by modifying the tone of influential statements, or enable fraudulent activities by creating emotionally persuasive or dis- tressing audio content. Additionally, by altering the emotional undertones of statements, EF could be used to damage reputations, potentially making neutral or positive statements appear hostile or inappropriate. The development and proliferation of EF technology under- score an urgent need for robust detection methods and ethical standards to mitigate the risks associated with emotion-altered audio. Detecting EF is particularly critical in sensitive domains such as security, legal pro- ceedings, and mental health care, where the integrity of emotional expression is essential for trust and ef- fective communication. Despite the advancements in audio deepfake detection (ADD) methods, which have greatly improved in identifying manipulated or syn- thetic audio, EF detection remains a largely unexplored field. This gap highlights the importance of developing dedicated research efforts and advanced methodologies capable of identifying emotion-based manipulations within audio content, addressing the unique challenges presented by emotion-modified deepfakes, and estab- lishing safeguards to prevent their misuse.

# LITERATURE REVIEW:

In recent years, audio deepfake detection (ADD) has become a critical area of study, especially as synthetic audio becomes increasingly realistic. Traditional ADD methods focus on identity-based manipulations, where speaker characteristics or spoken content are altered to produce deceptive yet convincing audio. Unlike typ- ical deepfakes, EmoFake alters emotional tone while preserving speaker identity and linguistic content, ne- cessitating specific detection approaches. Zhao et al. [1] introduced the EmoFake dataset, highlighting these challenges and underscoring the need for tailored de- tection methods. Recent studies show that Multilingual Speech Pre-trained Models (PTMs), which integrate cross-linguistic data, achieve state-of-the-art results in detecting multilingual deepfakes, as demonstrated by Chetia Phukan et al. [2]. Models like UniSpeech- SAT and WavLM leverage speaker-aware pre-training, which enhances their ability to distinguish emotion shifts through contrastive learning and augmentation techniques [ 3]. Cross-lingual self-supervised learning models, such as XLS-R and MMS, extend ADD’s reach into high- and low-resource languages by developing language-agnostic features. These models exhibit ro- bust generalization across diverse linguistic contexts, making them highly adaptable for emotion-based detec- tion tasks [4, 5]. Despite these advancements, dedicated methods for detecting emotion-altered audio remain limited. By examining multilingual SFMs for Emo- Fake Detection (EFD), this study fills a gap, showing that fusion of SFMs offers enhanced detection capa- bilities over individual models. This work sets a new benchmark for EFD by validating the effectiveness of multilingual SFMs in emotion-modified audio detec- tion.

# PROBLEM STATEMENT:

With advancements in synthetic audio technology, deepfakes have evolved beyond identity manipulation, giving rise to Emotion-based Audio Deepfakes or EmoFake. Unlike traditional deepfakes that alter a speaker's identity, EmoFake specifically modifies the emotional tone while retaining the speaker's voice and linguistic content, creating highly realistic and subtle manipulations. Such changes can lead to unintended emotional perceptions, affect social interactions, and even manipulate listeners’ psychological responses. This manipulation has serious implications in domains where emotional authenticity is crucial, such as media, legal proceedings, and interpersonal communication.

The complexity of detecting EmoFake lies in the subtlety of emotional manipulation, which often evades detection by conventional audio deepfake detection (ADD) methods that focus primarily on speaker identity changes. Consequently, there is an urgent need for specialized detection methods that can accurately identify emotion-altered audio content. Multilingual Speech Foundation Models (SFMs) have shown promise in addressing this gap due to their cross-linguistic pre-training, which helps capture emotional nuances across languages. However, the field lacks a robust framework for leveraging these models effectively in emotion-based deepfake detection tasks.

**OBJECTIVES:**

1. Develop a Robust Detection Framework for Emotion-Based Audio Manipulation: Design and implement a framework specifically tailored for detecting EmoFake instances, where only the emotional tone is altered while maintaining the speaker's identity and linguistic content.
2. Evaluate the Effectiveness of Multilingual Speech Foundation Models (SFMs): Perform a comparative analysis of state-of-the-art (SOTA) SFMs, including multilingual, monolingual, and speaker recognition models, to determine their efficacy in detecting emotion-altered audio across multiple languages.
3. Enhance Detection Accuracy through SFM Fusion: Propose a novel fusion strategy that combines the strengths of multiple SFMs to improve the accuracy and reliability of EmoFake detection beyond the capabilities of individual models.
4. Establish a New Benchmark for EmoFake Detection (EFD): Set a performance standard by comparing the proposed approach with existing methods in monolingual and cross-lingual contexts, validating the advantages of using fused multilingual SFMs.
5. Contribute to Ethical Standards and Security in Audio Deepfake Detection: Provide insights into the importance of detecting emotion-based audio manipulations, contributing to the development of ethical standards and safeguarding against misuse in fields where em**otional authenticity is critical.**

# Methodology:

* Data Preprocessing

The dataset for this study is based on the Emotional Speech Database (ESD), which includes recordings in English and Chinese, with samples spanning five primary emotions: Neutral, Happy, Angry, Sad, and Surprise. The EmoFake dataset simulates emotion-altered audio samples using seven open-source emotional voice conversion (EVC) models. These models employ techniques such as CycleGAN and StarGAN to modify emotional expressions while maintaining speaker identity and linguistic content. Preprocessing involves segmenting audio samples, resampling to match the requirements of specific SFMs, and normalizing audio features to ensure consistency across input data.

* Model Selection

We selected several SFMs based on their proven effectiveness in capturing nuanced speech characteristics. The models include UniSpeech-SAT, WavLM, wav2vec 2.0, x-vector, Whisper, XLS-R, and MMS. Each model brings unique strengths in handling multilingual data, diverse emotional expressions, and complex acoustic environments. These models are integrated into our framework to extract features that capture both linguistic and emotional nuances, essential for accurately identifying emotion manipulations.

* Feature Extraction

Each SFM processes the preprocessed audio data to extract high-dimensional embeddings. For instance, wav2vec 2.0 extracts features directly from raw audio using convolutional and Transformer layers, providing robust embeddings even in low-resource conditions. Similarly, WavLM and UniSpeech-SAT include techniques such as speaker-aware pre-training and contrastive learning to capture emotion-related information. The extracted features from each model are then processed using average pooling to create a standardized embedding representation, which is fed into the classifier for detection.

* Fusion of Multilingual SFMs

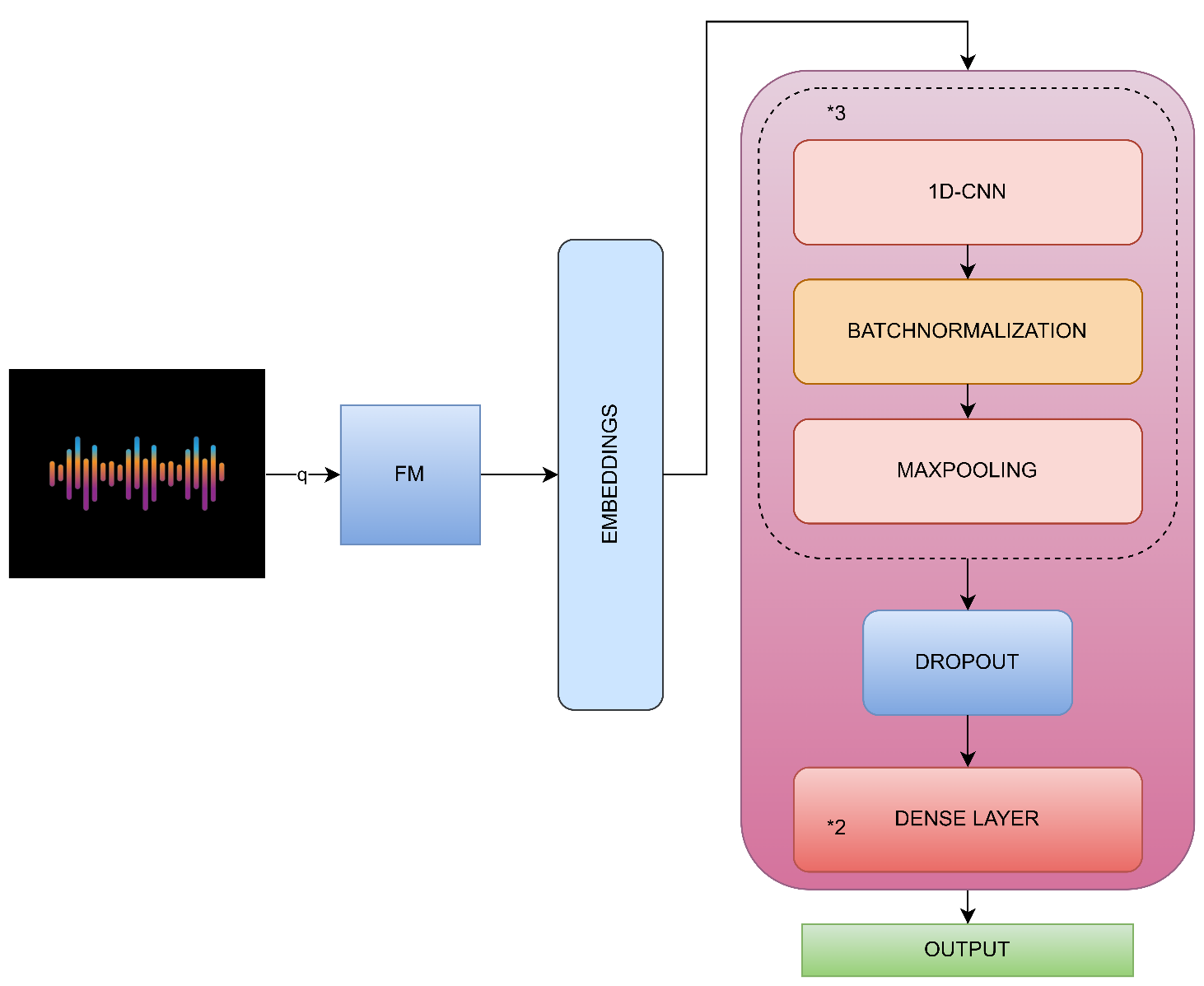
To enhance the accuracy and robustness of EmoFake detection, we propose a fusion approach that combines the outputs of multiple SFMs. By leveraging the complementary strengths of these models, the fusion approach captures a wider range of emotional and linguistic subtleties. We experimented with several fusion techniques, such as concatenating embeddings from different models and using ensemble methods to integrate model predictions. This fusion strategy enables the model to detect nuanced emotion alterations more effectively than individual SFMs.

* Model Training and Evaluation

We implement Convolutional Neural Network (CNN) and Fully Connected Network (FCN) architectures for classifying EmoFake samples. The CNN model consists of three 1D convolutional layers with increasing filter sizes (64, 128, 256), each followed by batch normalization, dropout (rate 0.2), and max-pooling layers to capture sequential features effectively. The final layer is a softmax-activated dense layer with two neurons for binary classification. The FCN, on the other hand, includes dense layers with 128, 128, and 64 neurons, each using ReLU activation, batch normalization, and dropout (rate 0.2 for initial layers, 0.5 for the final layer). Both models are trained for a maximum of 100 epochs using the Adam optimizer with an initial learning rate of 1e-3. We incorporate ReduceLROnPlateau to adapt the learning rate based on validation loss and EarlyStopping to halt training if no improvement is observed.

* Evaluation Metrics

The models are evaluated using the Equal Error Rate (EER), which provides an unbiased measure of the model's performance. EER scores are reported for both monolingual (English and Chinese) and cross-lingual (English-to-Chinese, Chinese-to-English) settings to assess the generalization capability of the fusion model. The experiments also compare individual SFM performance against the fused model to validate the benefits of the fusion approach.

****

1. **CONCLUSION:**

In this study, we address the task of Emotion-based Deepfake Detection (EFD), focusing on EmoFake (EF)—a type of audio deepfake where the original emotion is manipulated to express a different emotion. We hypothesize that multilingual speech foundation models (SFMs) are particularly effective for EFD due to their exposure to varied linguistic features such as pitch, tone, and intensity, acquired through pre-training across multiple languages. To test this, we conducted an extensive comparative analysis of state-of-the-art (SOTA) SFMs, including multilingual, monolingual, and speaker recognition models. Our results validated our hypothesis, confirming the superior performance of multilingual SFMs for EFD. Furthermore, we explored SFM fusion and developed a novel framework that leverages the complementary strengths of SFMs to enhance EFD accuracy. Our findings demonstrate that the fusion of multilingual SFMs significantly outperforms individual SFMs and baseline fusion methods, setting a new SOTA benchmark in EFD.

# REFERENCES:

[1] Y. Zhao, J. Yi, J. Tao, C. Wang, X. Zhang, and Y. Dong, “Emofake: An initial dataset for emotion fake audio detection,” arXiv preprint arXiv:2211.05363, 2022.

[2] O. Chetia Phukan, G. Kashyap, A. B. Buduru, and R. Sharma, “Heterogeneity over homogeneity: Investigating multilingual speech pre-trained models for detecting audio deepfake,” in Findings of the Association for Computational Linguistics: NAACL 2024. Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 2496–2506.

[3] S. Chen, Y. Wu, C. Wang, Z. Chen, Z. Chen, S. Liu, J. Wu, Y. Qian, F. Wei, J. Li, and X. Yu, “Unispeech-sat: Universal speech representation learning with speaker aware pre-training,” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6152–6156, 2021.

[4] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino et al., “Xls-r: Self-supervised cross-lingual speech representation learning at scale,” in Interspeech, 2022, pp. 2278–2282.

[5] V. Pratap, A. Tjandra, B. Shi, F. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi et al., “Scaling speech technology to 1,000+ languages,” Journal of Machine Learning Research, vol. 25, no. 97, pp. 1–52, 2024.

[6] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao et al., “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” IEEE Journal of Selected Topics in Signal Processing, vol. 16, no. 6, pp. 1505–1518, 2022. [7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in neural information processing systems, vol. 33, pp. 12 449–12 460, 2020.