OPEN ACCESS
VOL. 8, NO. 1, PP. 53–61, NOVEMBER 2025

ISSN : 2622-8106 (ONLINE)
DOI: 10.20895/INISTA.V8I1.2096

Journal of Informatics, Information System, Software Engineering and Applications (INISTA)

Multi-Layer Perceptron with Advanced Acoustic Features for Speech Emotion Recognition in Education Evaluation

Muhammad Afiq Tamamul Wafa

Departement of Informatics, Telkom University Purwokerto, 53147, Indonesia

*Corresponding email: wafanakha@student.telkomuniversity.ac.id

Received: January 12th, 2025; Revised: February 9th, 2026; Accepted: February 11th, 2026.

Abstract

Traditional methods for evaluating lecturer performance in education, such as student surveys, are often limited by their nature. This study explores the development of an objective, a framework to complement these evaluations through Speech Emotion Recognition (SER). This Research utilizes a specialized Indonesian speech emotion dataset, applying data augmentation techniques to enhance model generalization. A set of advanced acoustic features, including Mel Frequency Cepstral Coefficients (MFCCs), Chroma, and Spectral Contrast, along with their statistical variations, is used to create representations of the vocal expressions. A Multi Layer Perceptron (MLP) neural network was designed and trained on these features to classify five different emotions: happy, angry, sad, surprised, and neutral. The Research resulted in a model that demonstrated very good performance, achieving an overall classification accuracy of 94% with high precision, recall, and F1-scores across all emotions, indicating a balanced and reliable system. A critical feature analysis was also conducted, revealing the significance of the standard deviation of Chroma and MFCC features. This study shows that an MLP model paired with feature engineering can be used as a powerful and objective tool for providing deeper insights into student feedback, contributing a valuable new methodology for quality assurance in higher education.

Keywords: *Speech Emotion Recognition (SER), Deep Learning, Multi-Layer Perceptron (MLP), Lecturer Evaluation, Acoustic Feature Extraction, Higher Education.*

Corresponding Author:
*Muhammad Afiq Tamamul Wafa
Departement of Informatics, Telkom University Purwokerto
Jl. D.I. Panjaitan No. 128, Purwokerto 53147, Indonesia
Email: wafanakha@student.telkomuniversity.ac.id

I. INTRODUCTION

The continuous improvement of teaching quality is a cornerstone of excellence in higher education [1]. Central to this process is the evaluation of lecturer performance, such as student feedback surveys (known in Indonesia as EDOM Evaluasi Dosen oleh Mahasiswa). The reliability of these surveys has been a subject of debate, as they are easily influenced by various non-teaching-related biases, such as instructor gender, course difficulty, and grading leniency [2, 3]. The tone of voice, also known as vocal prosody, serves as a complex and often subconscious means of communication that conveys considerable emotional information, providing a possibility to be more objective [4]. The capability to evaluate this channel offers an opportunity to improve the fairness of lecturers' surveys.

Advancements in computing and deep learning have propelled the field of Speech Emotion Recognition (SER) into a viable technology for extracting meaningful emotional insights from audio data [5]. While bibliometric trends show a rapid growth in SER research [6], the majority of benchmark datasets have been concentrated on widely spoken languages like English, with a comparative lack of robust models tailored for specific linguistic and cultural contexts, such as the Indonesian language [7, 8].

This study aims to bridge this gap by designing and implementing a high-accuracy deep learning framework for SER, specifically applied to the context of lecturer evaluation in Indonesia. Utilizing the Indonesian speech emotion corpus for lecturer evaluation introduced by Rosita et al. [9], this research employs a sophisticated feature engineering strategy, unlike conventional methods, as its main highlight. We extract a comprehensive set of advanced acoustic features, including Mel-Frequency Cepstral Coefficients (MFCCs) [10], Chroma features [11], and Spectral Contrast to build a detailed representation of vocal emotional expression.

The core of our proposed framework is a Multi-Layer Perceptron (MLP), a powerful deep learning architecture, trained to classify emotions with high fidelity. This paper details the complete methodology, from data preprocessing and augmentation to model training and evaluation. We also demonstrate that this approach not only achieves a very good level of classification accuracy but also provides a critical analysis of the specific vocal features that are most indicative of emotional states. This study presents a reliable and objective tool that can significantly augment traditional evaluation methods, contributing to a better understanding of the student learning experience and advancing the application of SER in higher education quality assurance.

II. RESEARCH METHOD

This study followed a structured methodology consisting of five stages: (1) preparation and augmentation of data, (2) extraction of acoustic features, (3) design of the deep learning model, (4) model training and evaluation, and (5) critical feature analysis. The entire framework was implemented in Python using libraries such as Librosa for audio processin [12], Scikit-learn for data handling and feature analysis [13], and TensorFlow with Keras for building and training the deep learning model [14].

A. Dataset and Corpus

The primary data source for this research was the Indonesian speech emotion dataset developed by Rosita, Salsabila, and Pamungkas [9], specifically designed for the context of lecturer evaluation (EDOM). The complete corpus consists of 1,600 audio samples in .wav format, providing a robust foundation for training deep learning models.

The dataset is categorized into five discrete emotion classes, which are further grouped by sentiment for analysis:

1. Neutral: 400 samples
2. Positive: 491 samples, comprising 250 Happy (Bahagia) and 241 Surprised (Terkejut)
3. Neutral Emotion: 619 samples, comprising 337 Angry (Marah) and 282 Sad (Sedih)

Each recording has an average duration of 3–5 seconds, a temporal window selected as sufficient to capture the essential prosodic and emotional characteristics of speech without introducing excessive silence.

To ensure ecological validity and variance, the dataset was constructed from two distinct sources. First, direct recordings were obtained from lecturers with diverse backgrounds and teaching subjects. These sessions utilized wireless microphones to ensure high audio clarity, with lecturers reading from speech scripts corresponding to specific emotional states. Second, these were supplemented by publicly available recordings from actors, used with proper permission.

To minimize model bias, the dataset was designed to maximize speaker diversity and was regularly validated during the training process. Ethical considerations were strictly observed; no personal identifiers were included in the metadata, ensuring the preservation of privacy for all subjects.

B. Data preprocessing and augmentation

Before feature extraction, all audio files were resampled to a uniform sampling rate of 22,050 Hz and truncated or padded to a consistent length of five seconds. This standardization is important for ensuring that all input features have a uniform dimension.

To enhance model generalization and mitigate the risk of overfitting, data augmentation techniques were applied [15]. For each original audio file, three augmented versions were generated:

1. Noise Injection: White noise with a small amplitude was added to the signal. This was implemented by generating a noise vector and adding it to the normalized signal, scaled by a small amplitude factor to improve the model's ability to reduce background interference.
2. Pitch Shifting: The pitch of the audio was randomly shifted up or down by a small interval (± 2 semitones) using `librosa.effects.pitch_shift` to help the model learn pitch-invariant features.
3. Time Stretching: The duration of the audio was randomly stretched or compressed by a factor between 0.8 and 1.2 using `librosa.effects.time_stretch`, making the model less sensitive to variations in speech rate.

C. acoustic feature extraction

To create a rich, low-dimensional representation of the vocal expressions, a comprehensive set of advanced acoustic features was extracted from each audio clip. This process yielded a 64-dimensional feature vector for each sample, which was constructed by concatenating the mean and standard deviation for three distinct feature types. The specific librosa functions and configurations used for this extraction are detailed in Table I

The 64-dimensional feature vector was composed as follows:

1. Mel-Frequency Cepstral Coefficients (MFCCs): 13 MFCCs were extracted to represent the timbral quality of the voice [10]. The mean and standard deviation of these coefficients over the five-second window were calculated, resulting in 26 features.
2. Chroma Features: 12 Chroma features were extracted to capture the pitch and harmonic content of the speech [11]. The mean and standard deviation were computed, adding another 24 features.
3. Spectral Contrast: 7 spectral contrast features were extracted to measure the relative distribution of energy across the frequency spectrum. The mean and standard deviation were calculated, contributing to the final 14 features.

The standard deviation was included for each feature set to explicitly model the dynamic variation and modulation of vocal characteristics, which is highly indicative of emotional arousal.

TABLE I. ACOUSTIC FEATURE EXTRACTION PARAMETERS

Feature	Librosa function	Parameter	Value
		sr	22050
Shared	STFT	`n_fft`	2048
		`hop_length`	512
MFCC	`librosa.feature.mfcc`	`n_mfcc`	13
Chroma	`librosa.feature.chroma_stft`	`n_chroma`	12
Spectral Contrast	`librosa.feature.spectral_contrast`	`n_bands`	7

D. Model architecture: Multi-Layer Perceptron (MLP)

Multi-Layer Perceptron (MLP), a type of feedforward deep learning neural network, was designed using the Keras Sequential API to perform the emotion classification task. The architecture was specifically built to balance complexity and performance, consisting of an input layer, three hidden layers, and an output layer.

The model architecture is detailed as follows:

1. Input Layer: A layer with 512 neurons, accepting the 64-dimensional feature vector.
2. Hidden Layer 1: A layer with 256 neurons.
3. Hidden Layer 2: A layer with 128 neurons.
4. Output Layer: A layer with 5 neurons, corresponding to the five emotion classes.

The model accepts the 64-dimensional feature vector as input. All hidden layers utilised the Rectified Linear Unit (ReLU) activation function [16]. To prevent overfitting and to stabilize the training process, Batch Normalisation and Dropout layers (with a rate of 0.5) were applied after each hidden layer [17]. The final output layer used a Softmax activation function to produce a probability distribution over the five emotion classes.

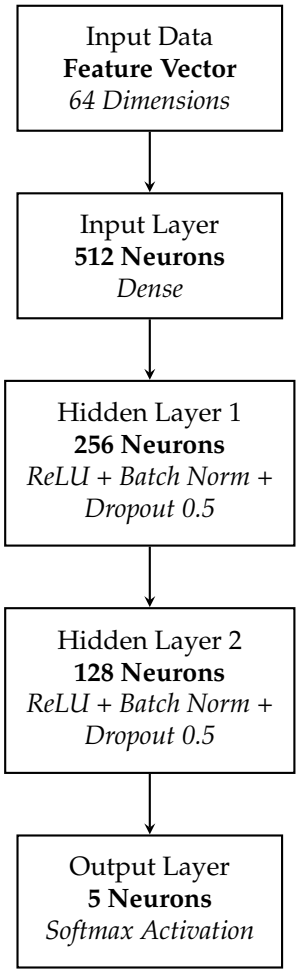

Fig. 1. Architecture of the MLP classifier

E. Training and Evaluation Protocol

The augmented and feature-extracted dataset was split into a training set (80%) and a testing set (20%) using a stratified sampling method to maintain the same class distribution in both sets. The features were then standardized using Scikit-learn's `StandardScaler`.

The MLP model was compiled with `CategoricalCrossentropy` and trained using the Adam optimiser [18] with an initial learning rate of 0.0001. To handle the class imbalance, class weights were calculated and applied during training using Scikit-learn's `compute_class_weights` with the parameter `class_weight="balanced"`, penalizing the model more for mistaking emotions.

Two callbacks were used to manage the training process: `EarlyStopping` to halt training if the validation loss did not improve for 20 consecutive epochs, and `ReduceLROnPlateau` to decrease the learning rate if training stagnates.

The model's performance was evaluated on the unseen test set using classification metrics: Accuracy, Precision, Recall, and F1-Score. A Classification Report and a Confusion Matrix were generated to provide a detailed, per-class analysis of the model's predictive performance.

TABLE II. MODEL TRAINING HYPERPARAMETER CONFIGURATION

Category	Parameter	Value
Optimizer	Optimizer	`Adam`
	Learning Rate	0.0001
Loss	Loss Function	`CategoricalCrossentropy`
	Class Weight	`Balanced`
Callbacks	`EarlyStopping`: monitor	`val_loss`
	`EarlyStopping`: patience	20
	`ReduceLROnPlateau`: monitor	`val_loss`
	`ReduceLROnPlateau`: factor	0.1
	`ReduceLROnPlateau`: patience	10

F. Feature Importance Analysis

For a feature Importance analysis, a Random Forest classifier was trained on the same dataset [19]. While the MLP is our primary model for predictive, the Random Forest algorithm offers a mechanism to calculate feature importance. This analysis was conducted after MLP training and evaluation to identify and rank the 64 acoustic features based on their contribution to the classification task, providing objective insights into which vocal patterns were most significant for distinguishing emotions.

TABLE III. RANDOM FOREST HYPERPARAMETERS FOR FEATURE IMPORTANCE

Parameter	Value
`n_estimators`	200
`random_state`	42
`class_weight`	`balanced`
`n_jobs`	-1
`max_depth`	`None`

III. RESULT

A. Model Performance ON Emotion Classification

After being trained on an augmented dataset of advanced acoustic features, the MLP model demonstrated a high degree of accuracy and reliability on the unseen test set. The model achieved an overall classification accuracy of 94.39%.

The results show high values for precision, recall, and F1-score in all categories, as shown in the Classification report Table IV. This indicates the model is well-balanced.

TABLE IV. DETAILED CLASSIFICATION REPORT

	Precision	***recall***	***f1-score***	***support***
Happy	0.94	0.91	0.92	160
Angry	0.97	0.93	0.95	234
Neutral	0.93	0.96	0.95	240
Sad	0.95	0.96	0.96	190
Surprised	0.91	0.96	0.93	120
Accuracy			0.94	944
Macro avg	0.94	0.94	0.94	0.94
Weighted avg	0.94	0.94	0.94	0.94

B. Confusion Matrix

A confusion matrix was used (Figure 2) to visualize the model's performance. The matrix shows the variable accuracy of model predictions.

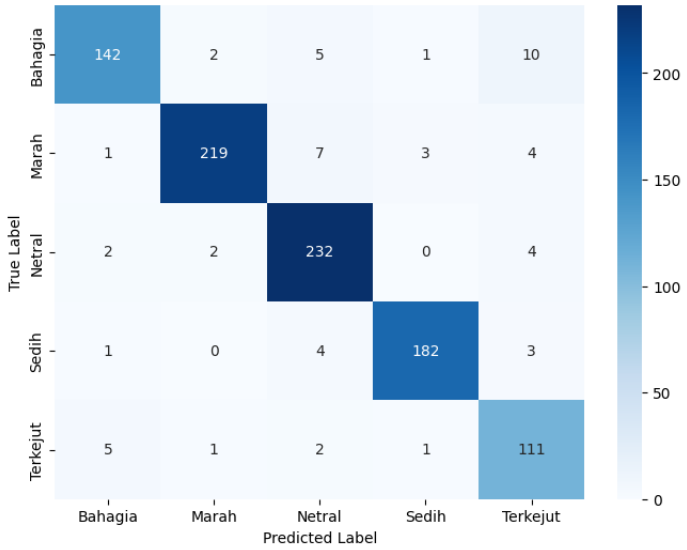

Fig. 2. Confusion Matrix of the Model

C. *Feature Importance*

The feature importance analysis, executed using a Random Forest model, identified the specific acoustic features that were most influential in the classification process. The results are visualized in Figure 3, which reveals that the standard deviation of various features was more significant than their mean values.

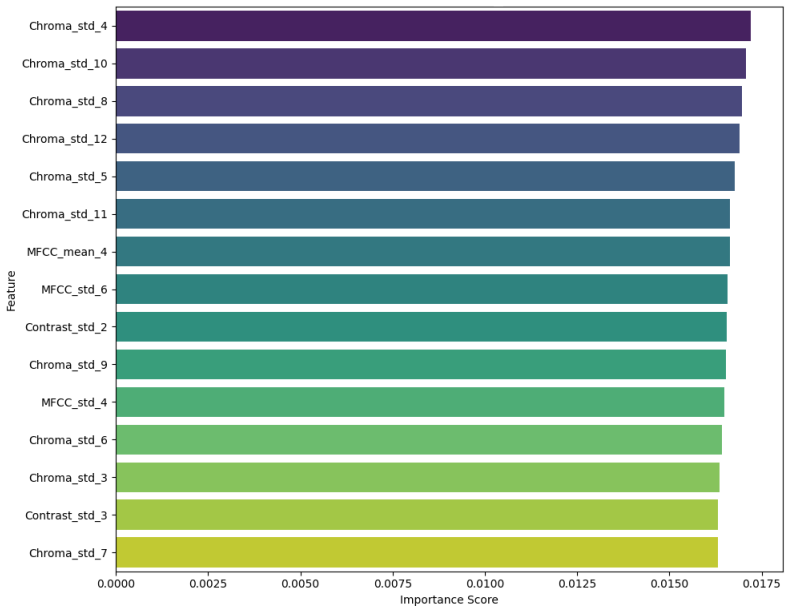

Fig. 3. Feature Importance Analysis

IV. DISCUSSION

This study successfully demonstrates the high potential of a deep learning based Speech Emotion Recognition (SER) framework for enhancing lecturer evaluation in higher education. The result of 94% accuracy indicates the development of a reliable system for capturing the emotional content of student feedback.

A. Interpretation of Findings

The performance of the Multi-Layer Perceptron (MLP) model can be attributed to two primary factors: the sophisticated feature engineering strategy of the deep learning. By extracting the standard deviation, not just the mean of MFCCs, Chroma, and Spectral Contrast, this study gives the model a dynamic representation of the audio features. The feature importance analysis (Figure 3) confirms this hypothesis, revealing that the model relied on the standard deviation of these features (e.g., `Chroma_std`, `MFCC_std`). This finding is highly significant and aligns with established theories in science, which advance that emotional expression is carried more through the dynamics of the voice, such as changes in pitch contour and vocal timbre, than through static, averaged acoustic properties [4].

The success of the MLP, a straightforward deep learning model with a strong feature set, makes performance achievable without relying on more complex architectures. This supports the notion that for certain SER tasks, the quality of feature representation can be as critical as the complexity of the model itself. The balanced performance across all five emotion classes further proves the effectiveness of combining data augmentation with class weighting, a crucial step for addressing the well-documented issue of class imbalance in emotion datasets [20].

B. Implication for Higher Education in Indonesia

The practical implications of this research for the method of quality assurance in Indonesian higher education are important. The developed method can serve as a powerful objective complement to the existing EDOM system. By analyzing the emotional content of verbal feedback, university administrators and lecturers can gain a deeper, more authentic understanding of the student experience.

This tool is not intended to replace subjective surveys but to enrich them. It can provide context to written comments and numerical ratings, helping to answer the "why" behind student satisfaction or dissatisfaction.

C. Limitations and Future Work

Despite the strong results, this study has several limitations that open avenues for future research. First, while the dataset improves ecological validity by including direct recordings from lecturers using wireless microphones, the speech samples both from lecturers and the supplementary actors are primarily scripted. Models trained on scripted or acted data may not always generalize perfectly to spontaneous, real-world student feedback, where emotions are often more subtle, co-articulated, and less prototypical [21]. Future work should focus on collecting spontaneous classroom recordings to test and refine the model's performance in a fully naturalistic environment.

Second, while the advanced acoustic features were highly effective, the feature extraction process is static. Exploring end-to-end deep learning models, such as those based on Transformers that can learn directly from raw spectrograms, could potentially capture even more complex temporal dependencies in the speech signal [22].

Finally, this study was only done in the Indonesian language. Expanding the research to include other languages or conducting cross-linguistic comparative studies could give us valuable insights into the universality of vocal emotional expressions, as evidence suggests that while some prosodic cues are universal, others are culture-dependent [23].

V. CONCLUSION

This research successfully addresses the critical need for objective mechanisms in higher education quality assurance by validating a deep learning framework designed to analyze the emotional content of student feedback. By responding to the limitations of traditional, subjective surveys, this study demonstrates that Speech Emotion Recognition can effectively serve as a reliable, data-driven complement to existing lecturer evaluation systems.

The methodology solves the problem of classifying complex vocal expressions by integrating a specialized feature engineering strategy with a Multi-Layer Perceptron architecture. Through the application of rigorous data augmentation and class-weighting techniques, the proposed model overcomes common challenges such as overfitting and data imbalance, resulting in a system capable of distinguishing between distinct emotional states with high consistency and balance.

The distinct novelty of this study lies in two key areas. First, it bridges a significant gap in the literature by developing a model specifically tailored to the Indonesian language within an educational context,

ensuring ecological validity where generalized models often fail. Second, and most critically, the research uncovers that the dynamic variation of vocal features represented by their standard deviation serves as a more potent predictor of emotion than static average values. This insight shifts the understanding of how vocal prosody should be analyzed for emotion recognition, establishing that the fluctuation of spectral and tonal characteristics is the primary carrier of emotional intent. Ultimately, this work provides a validated methodology for university administrators to gain a deeper, more authentic understanding of the student learning experience.

ACKNOWLEDGMENTS

Special thanks to Y. D. Rosita, Z. Salsabila, and A. R. P. Pamungkas for developing and making their specialized Indonesian speech emotion dataset available, which was fundamental to this research.

REFERENCES

[1] L. Harvey and D. Green, "Defining quality," *Assessment & Evaluation in Higher Education*, vol. 18, no. 1, pp. 9–34, 1993.

[2] A. Boring, K. Ottoboni, and P. B. Stark, "Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors," *ScienceOpen Research*, 2016.

[3] M. Braga, M. Paccagnella, and M. Pellizzari, "Evaluating students evaluations of professors," *Economics of Education Review*, vol. 41, pp. 71–88, 2014.

[4] K. R. Scherer, "Vocal affect expression: A review and a model for future research," *Psychological Bulletin*, vol. 99, no. 2, pp. 143–165, 1986.

[5] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[6] Y. D. Rosita, M. R. Firmansyah, and A. Utami, "Exploring bibliometric trends in speech emotion recognition (2020-2024)," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 14, no. 4, pp. 3421–3434, 2025.

[7] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[8] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[9] Y. D. Rosita, Z. Salsabila, and A. R. P. Pamungkas, "Lecturer evaluation from the perspective of speech emotion recognition with deep learning," in *2025 International Conference on Data Science and Its Applications (ICODSA)*, (Jakarta, Indonesia), pp. 565–571, 2025.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[11] M. Müller and S. Ewert, "Chroma toolbox: Matlab implementations for feature extraction in music audio signal processing," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, pp. 215–220, 2011.

[12] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, vol. 8, pp. 18–25, 2015.

[13] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[15] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[16] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 807–814, 2010.

[17] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[18] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[20] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 2, pp. 99–117, 2012.

[21] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[23] M. D. Pell, L. Monetta, S. Paulmann, and S. A. Kotz, "Recognizing emotions in a foreign language," *Journal of Nonverbal Behavior*, vol. 33, no. 2, pp. 107–120, 2009.