Proposed Solution for Voice-Based Age, Gender, and Emotion Recognition

Nguyễn Minh Nhựt - SE184534, Nguyễn Thành Trung - SE180355,
Trịnh Đình Hùng - SE173408, Nguyễn Hứa Hiệp - SE183787
Group 3, AI1804, FPT University, Ho Chi Minh City Campus

I. Proposed Solution

The proposed model for voice-based age, gender, and emotion recognition involves a comprehensive feature extraction process. Key features include Mel-Frequency Cepstral Coefficients (MFCC), Delta Mel-Frequency Cepstral Coefficients (delta-MFCC), Delta delta Mel-Frequency Cepstral Coefficients (delta delta-MFCC), Pitch, Filter-Bank Energies, Zero-Crossing Rate (ZCR), and ZCR Density. After that, we are using Principal component analysis (PCA) to reduce feature. In this research, we develop three separate classification models: one for age recognition, one for gender recognition, and one for emotion recognition. We utilize various architectures for classification and recognition, including Support Vector Machine (SVM), Long - Short Term Memory (LSTM), and Convolutional Neural Network (CNN)-based architectures such as DummyNet 1D, RezoNet, and ExpoNet. The architecture of model in Figure 1.

II. Performance Metrics

In this research, we utilize precision, recall, and F1-score as metrics for evaluating the performance of our classification models.

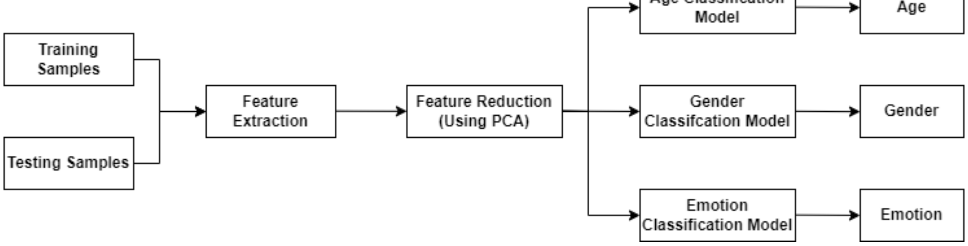
- **Precision:** Precision evaluates the percentage of accurately predicted positive instances among all instances identified as positive by the model. This metric is determined by dividing the number of true positives by the total of true positives and false positives. It is followed by Eq. 1.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1}$$

- **Recall:** Recall, known as sensitivity, evaluates the proportion of correctly predicted positive instances out of all true positive instances within the dataset. In Eq. 2, this metric is calculated by dividing the number of true positives by the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{2}$$

- **F1-score:** The F1-score, depicted in Eq. 3, represents the harmonic mean of precision and recall. This metric offers a balanced assessment by considering both precision and recall, rendering it especially valuable in scenarios where the dataset exhibits imbalance.



Hình 1. Voice-Based Age, Gender, and Emotion Recognition Model Architecture

Sound processing	**Feature**
MFCC	13
Delta MFCC	13
Delta delta MFCC	13
Picth	1
Filter - Bank Energies	26
Zero-Crossing Rate	1
ZCR Desity	1
Total	**68**

Bảng I
CAPTION

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

- **Macro average:** Macro average is used to describe a specific type of averaging technique utilized in multi-class classification tasks. It is commonly applied to compute performance metrics like precision in Eq. 4, recall in Eq. 6, or F1 score in Eq. 5, particularly in situations involving imbalanced datasets. In the case of the Macro average, the metric is computed individually for each class, and subsequently, the average is calculated across all classes.

$$\text{Macro Precision} = \frac{\text{Precision of Class 1} + \text{Precision of Class 2} + \ldots + \text{Precision of Class N}}{N} \tag{4}$$

$$\text{Macro F1-score} = \frac{2 \times (\text{Macro Precision} \times \text{Macro Recall})}{\text{Macro Precision} + \text{Macro Recall}} \tag{5}$$

$$\text{Macro Recall} = \frac{\text{Recall of Class 1} + \text{Recall of Class 2} + \ldots + \text{Recall of Class N}}{N} \tag{6}$$

- **Weighted average:** The weighted precision, recall, and F1-score functions are essential metrics in evaluating the performance of multi-class classification models. They incorporate weights assigned to each class to calculate a weighted average, reflecting the significance or contribution of individual classes to overall performance. Weighted precision assesses the accuracy of positive predictions, weighted recall measures the completeness of positive predictions, precision in Eq. 7 and balances recall in Eq. 8 and weighted F1-score is presented in Eq. 9, providing a comprehensive evaluation of model effectiveness while considering class-specific importance.

$$\text{Weighted Precision} = \frac{p_1 \times w_1 + p_2 \times w_2 + \ldots + p_N \times w_N}{\text{Total Weight}} \tag{7}$$

where p_i is the precision of class i and w_N is the weight of class i.

$$\text{Weighted Recall} = \frac{r_1 \times w_1 + r_2 \times w_2 + \ldots + r_N \times w_N}{\text{Total Weight}} \tag{8}$$

where r_N is the recall of class i and w_N is the weight of class i.

$$\text{Weighted F1 - Score} = \frac{f_1 \times w_1 + f_2 \times w_2 + \ldots + f_N \times w_N}{\text{Total Weight}} \tag{9}$$

where f_N is the F1-Score of class i and w_N is the weight of class i.

III. DATASET

For Age and Gender model, we use dataset from Common Voice Mozilla [1] for Age and Gender Classification Model, and RAVDESS dataset [2], CREMA-D dataset [3], TESS dataset [4], SAVEE dataset [5], distribute in Table V, for Emotion Classification Model. The age, gender, and sentiment distributions in the dataset are shown in the tables II, III, IV.

Class	Percents
0 - 19	6%
20 - 29	25%
30 - 39	14%
40 - 49	9%
50 - 59	5%
60 - 69	4%
70 - 79	1%
No information	36%

Bảng II
COMMON VOICE CORPUS 17.0, AGE ANALYSIS

Class	Percents
Female	19%
Male	53%
No information	27%

Bảng III
COMMON VOICE CORPUS 17.0, GENDER ANALYSIS

IV. RELATED WORK

A. Paper

Noushin Hajarolasvadi et al. [6] propose a new system to recognize emotions from speech. It works by first breaking down speech recordings into short segments and extracting features like pitch and intensity. Then, it uses a clustering technique to identify the most important segments and creates a 3D representation based on those segments. Finally, a special kind of neural network analyzes these 3D representations to identify emotions. Experiments show this system performs better than previous methods.

Dr. Madhu M. Nashipudimath et al. [7] tackle recognizing both feelings and a person's sex just by listening to their voice. Men and women naturally have different vocal qualities, and these along with various emotions can be identified through computer analysis. To achieve this, the system first cleanses the speech recording and extracts key characteristics like MFCCs. Then it utilizes a special kind of classifier to categorize the emotions and gender. This approach is reported to be more effective than past methods, making it potentially useful in healthcare, virtual assistants, security systems and other areas where machines interact with people.

Nammous, Mohammad K. and et al. [8] explore using LSTMs (a powerful type of neural network) for speaker recognition, specifically focusing on situations with limited training data. Traditionally, a lot of training data is

Class	Percents
Angry	16.7%
Happy	16.46%
Sad	16.35%
Neutral	14.26%
Fearful	16.46%
Disgusted	15.03%
Surprised	4.74%

Bảng IV
EMOTION ANALYSIS

Dataset Name	Percents
CREMA-D	58.15%
TESS	21.88%
RAVDESS	16.22%
SAVEE	3.75%

Bảng V
EMOTION DATASET ANALYSIS

needed for good results in speaker recognition. This paper shows that LSTMs can achieve high accuracy even with a smaller training set. The paper compares LSTMs to a different kind of neural network (feed-forward) and finds LSTMs perform better for speaker recognition, gender identification, and language identification – even when training data is limited.

B. Code

Yousef Kotp et al.'s notebook [9] is all about experimenting different set of features for audio representation and different CNN-based architectures for building speech emotion recognition model. The notebook contains implementation for three new CNN-based architectures for speech emotion recognition. which are: DummyNet 1D, RezoNet, ExpoNet.

Voice gender recognition from SuperKogito [10] can be achieved by analyzing speech patterns. Here, researchers use a dataset of speakers with known genders. They then extract Mel-frequency cepstral coefficients (MFCCs) which convert voice samples into a format that highlights speaker characteristics. Next, they train Gaussian mixture models (GMMs) to represent the male and female speaker distributions based on these MFCCs. Finally, when analyzing an unknown voice sample, they can compare its MFCCs to the trained GMMs and predict the speaker's gender based on which model shows a better fit. The system results in a 95% accuracy of gender detection.

TÀI LIỆU

[1] "Mozilla voice dataset." [Online]. Available: https://commonvoice.mozilla.org

[2] S. R. Livingstone and F. 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*, vol. 13, no. 5, p. e0196391, 2018.

[3] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[4] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[5] K. Dupuis and M. K. Pichora-Fuller, *Toronto emotional speech set (TESS)*. University of Toronto, Psychology Department, 2010.

[6] N. Hajarolasvadi and H. Demirel, "3d cnn-based speech emotion recognition using k-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, p. 479, 2019.

[7] M. M. Nashipudimath, P. Pillai, A. Subramanian, V. Nair, and S. Khalife, "Voice feature extraction for gender and emotion recognition," in *ITM Web of Conferences*, vol. 40. EDP Sciences, 2021, p. 03008.

[8] M. K. Nammous and K. Saeed, "Natural language processing: Speaker, language, and gender identification with lstm," *Advanced Computing and Systems for Security: Volume Eight*, pp. 143–156, 2019.

[9] "Speech emotion recognition." [Online]. Available: https://github.com/yousefkotp/Speech-Emotion-Recognition

[10] "Voice based gender recognition." [Online]. Available: https://github.com/SuperKogito/Voice-based-gender-recognition