Speech Region Recognition and Speech Gender Recognition Using Deep Learning on Vietnamese Accent Dataset

1st Le Nhat Truong
CE181823
FPT University
Tra Vinh, Viet Nam
lenhattruongtv04@gmail.com

2nd Le Thanh Thao
CE181385
FPT University
Kien Giang, Viet Nam
lethanhthao112.cxm@gmail.com

3rd Huynh Minh Thy
CE181186
FPT University
Can Tho, Viet Nam
thylun2k4@gmail.com

4th Nguyen Vo An Xuan
CE180467
FPT University
An Giang, Viet Nam
anxuan4553@gmail.com

5th Nguyen Trung Truc
CE181944
FPT University
Bac Lieu, Viet Nam
trucntce181944@fpt.edu.vn

***Abstract*—This study presents a deep learning-based approach for Speech Region Recognition (SRR) and Speech Gender Recognition (SGR) using the Vietnamese Accent 3 Regions (VA3R) dataset. The VA3R dataset, collected from YouTube, contains 8,254 audio samples representing three Vietnamese regions (North, Central, South) and two genders (Male, Female). We employ a 2D Convolutional Neural Network (CNN2D) architecture to classify regional accents and genders, achieving high accuracy through rigorous data preprocessing, feature extraction, data augmentation, and hyperparameter optimization using Keras Tuner. The proposed system preprocesses audio with loudness normalization, extracts features such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma, and spectral features, and balances the dataset to address class imbalances. The CNN2D models for SRR and SGR achieve test accuracies of 99.66% and 99.66%, respectively, with significantly reduced model sizes after optimization (from 6.74 MB to 2.89 MB for SRR and 1.07 MB for SGR). These results demonstrate the effectiveness of the proposed approach in accurately classifying regional accents and genders in Vietnamese speech, with potential applications in speaker profiling, forensic analysis, and personalized human-computer interaction.**

***Index Terms*—speech region recognition, speech gender recognition, deep learning, Vietnamese accent, CNN2D, VA3R dataset, feature extraction, hyperparameter tuning**

I. Introduction

Speech Region Recognition (SRR) and Speech Gender Recognition (SGR) are critical tasks in speech processing, enabling systems to identify the regional accent and gender of a speaker from audio samples. These tasks have significant applications in speaker profiling, forensic analysis, personalized human-computer interaction, and cultural studies. In the context of Vietnamese speech, regional accents (North, Central, South) exhibit distinct phonetic and prosodic characteristics, while gender differences manifest in pitch, tone, and vocal patterns. Accurately classifying these attributes requires robust models capable of capturing intricate speech features.

Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have shown remarkable success in speech-related tasks due to their ability to learn hierarchical features from raw audio data. This study focuses on developing a deep learning-based system for SRR and SGR using the Vietnamese Accent 3 Regions (VA3R) dataset, a novel dataset collected by our team from YouTube. The dataset comprises 8,254 three-second audio samples, representing three Vietnamese regions (North: 2,857, Central: 2,853, South: 2,540) and two genders (Male: 4,761, Female: 3,489).

Our approach leverages a 2D Convolutional Neural Network (CNN2D) architecture, optimized for speech classification tasks. We address challenges such as class imbalance, audio loudness variability, and computational efficiency through rigorous data preprocessing, feature extraction, data augmentation, and hyperparameter tuning using Keras Tuner. The proposed system achieves high classification accuracy while maintaining low computational complexity, making it suitable for real-time applications.

The main contributions of this study are:

- Development of the VA3R dataset, a comprehensive collection of Vietnamese speech samples for SRR and SGR.
- A CNN2D-based approach for accurate classification of regional accents and genders, achieving test accuracies of 99.66% for both tasks.
- Optimization of model size and inference speed through hyperparameter tuning, reducing model parameters significantly while maintaining performance.
- A preprocessing pipeline that ensures consistent audio input for real-time applications, addressing loudness vari-

This work was supported by FPT University. The authors would like to thank the university for providing resources and support for this research.

ability and class imbalance.

The rest of the paper is organized as follows: Section II describes the dataset, preprocessing, and model development. Section III presents the training results and evaluation metrics. Section IV discusses the findings and limitations, and Section VI concludes the study with future directions.

II. Methodology

A. Data Collection Process

To construct the Vietnamese Accent 3 Regions (VA3R) dataset, we developed a systematic data collection pipeline utilizing YouTube as the primary source of audio data. The process involved several steps to ensure the dataset captures authentic Vietnamese speech across the three regions (North, Central, South) and includes gender diversity (Male, Female). The detailed steps are as follows:

1) **Video Selection and Link Collection:** We manually identified and collected links to YouTube videos containing spoken Vietnamese content from the three target regions: North (Bac Ninh), Central (Hue), and South. These videos were sourced from diverse channels, including interviews, vlogs, and regional media, to ensure a wide range of speakers and speaking styles. A total of 100 videos were selected.
2) **Audio Download Using yt_dlp:** The audio tracks from the selected YouTube videos were downloaded using the `yt_dlp` library [1], a Python-based tool for extracting audio and video content from online platforms. We configured `yt_dlp` to extract audio in WAV format at a sampling rate of 48 kHz to maintain high fidelity, resulting in a total of 10 hours of raw audio data.
3) **Voice Activity Detection (VAD):** To isolate segments containing human speech, we employed the `pyannote/voice-activity-detection` model from Hugging Face [2], a pre-trained model designed for voice activity detection. This model uses a deep learning-based approach to distinguish between speech and non-speech segments, such as background noise or music. The raw audio was segmented into clips containing human speech, yielding approximately 6 hours of speech data after filtering out non-speech segments.
4) **Manual Gender Labeling:** The speech segments were manually labeled for gender (Male or Female) by our team. This step involved listening to each segment and assigning a gender label based on the perceived voice characteristics, such as pitch and tone. To ensure accuracy, each segment was reviewed by at least two team members, and discrepancies were resolved through discussion. This process resulted in 4,761 male samples and 3,489 female samples.
5) **Regional Labeling and Segmentation into 3-Second Clips:** The speech segments were further labeled according to their region of origin (North, Central, South) based on the metadata of the source videos. We then developed a Python script, `cut_3s.py`, to segment the audio into 3-second clips. This script iterates through each audio file, extracts 3-second segments with a sliding window approach (with a 1-second overlap to maximize data usage), and saves them as individual WAV files. Segments shorter than 3 seconds were padded with zeros, while longer segments were trimmed accordingly. This process produced 8,254 3-second audio clips, forming the final VA3R dataset.
6) **File Naming Convention:** Each audio file in the VA3R dataset is named in the format `{a}-{b}-{c}-{d}.wav`, where:
   - `a`: Region (North: 1, Central: 2, South: 3).
   - `b`: Sound type (1: no noise, 2: noise, 3: loud background music).
   - `c`: Gender (0: male, 1: female).
   - `d`: Ordinal number.

   This naming convention facilitates efficient parsing and organization of the dataset for subsequent preprocessing and model training.

The resulting VA3R dataset provides a robust foundation for training SRR and SGR models, capturing the diversity of Vietnamese regional accents and gender-specific vocal characteristics. The dataset is publicly available on Kaggle [3] for further research and reproducibility.

B. Dataset Description

We utilize the Vietnamese Accent 3 Regions (VA3R) dataset, collected by our team and uploaded to Kaggle on March 22, 2025. The dataset contains 8,254 three-second audio samples from three regions in Vietnam: North (Bac Ninh, 2,857 samples), Central (Hue, 2,853 samples), and South (2,540 samples). It also includes gender labels: 4,761 male samples and 3,489 female samples. The distribution of samples by region and gender is shown in Figures 1, 2, 3, and Table I.

Region_Gender	Counts
North_Male	2128
North_Female	729
Central_Male	1480
Central_Female	1373
South_Male	1153
South_Female	1387

TABLE I: Counts by Region and Gender

Each audio file is named in the format `{a}-{b}-{c}-{d}.wav`, where:

- `a`: Region (North: 1, Central: 2, South: 3).
- `b`: Sound type (1: no noise, 2: noise, 3: loud background music).
- `c`: Gender (0: male, 1: female).
- `d`: Ordinal number.

C. Data Preprocessing and Feature Extraction

1) Feature Extraction: We extract features using techniques similar to those applied in Speech Emotion Recognition (SER). The following features are computed using Librosa:

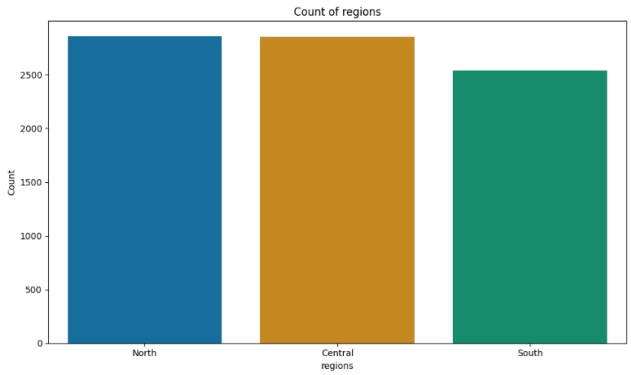


Fig. 1: Count the number of each region

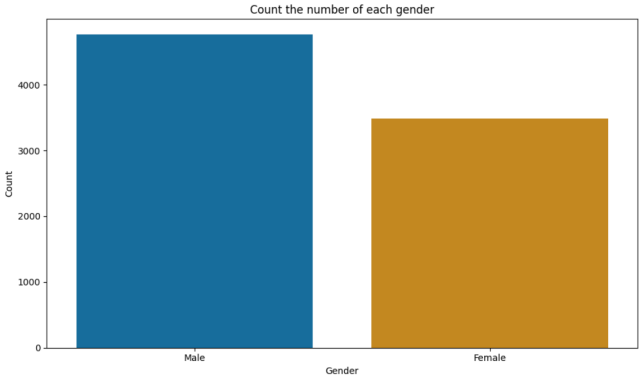


Fig. 2: Count the number of each gender

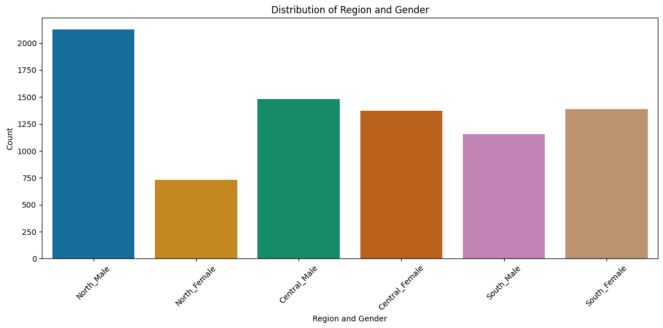


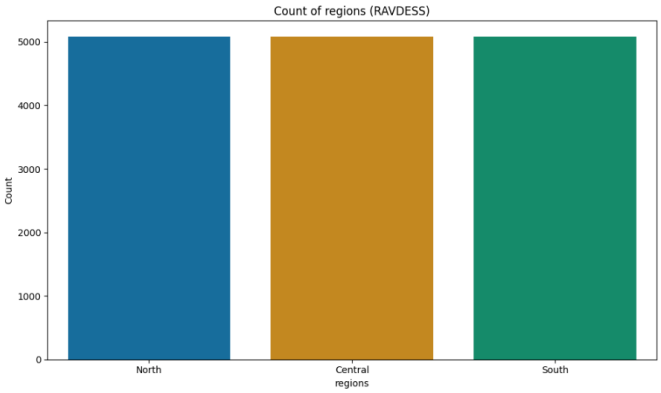
Fig. 3: Region and Gender

- **MFCCs (Mel-Frequency Cepstral Coefficients):** 8 coefficients are extracted to capture spectral characteristics while reducing computational complexity.
- **Chroma Features:** 12 features representing pitch classes (C, C#, ..., B) to capture harmonic content.
- **Spectral Centroid:** Indicates the center of mass of the spectrum, reflecting sound brightness.
- **Spectral Bandwidth:** Measures the width of the spectral range, distinguishing voice qualities.

Audio signals are standardized to a fixed duration of 3 seconds using the `pad_or_trim_audio()` function, which pads shorter samples with zeros or trims longer ones.

2) Data Augmentation and Class Balancing: The VA3R dataset exhibits class imbalances (e.g., fewer female samples in the North). We split the dataset into two subsets: one for SRR (by region) and one for SGR (by gender). To balance the dataset, we apply data augmentation by multiplying the class with the least samples by 2 using noise addition (scale 0.02). Other classes are also augmented to ensure the total number of samples per class does not exceed twice the smallest class. After augmentation, we obtain 15,240 samples for SRR and 13,956 samples for SGR, as shown in Figures 4 and 5.

```
region Counts:
North: 5080
Central: 5080
South: 5080
```




```
gender Counts:
Male: 6978
Female: 6978
```

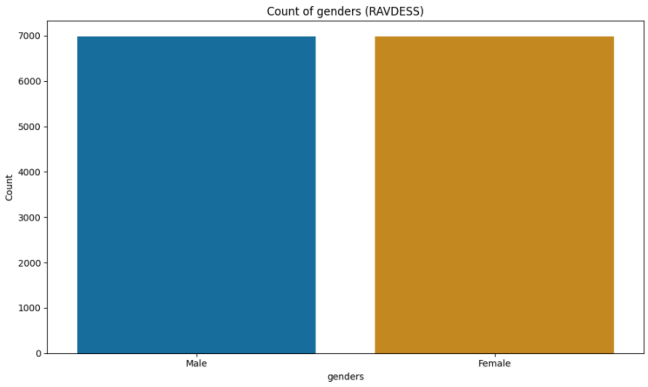
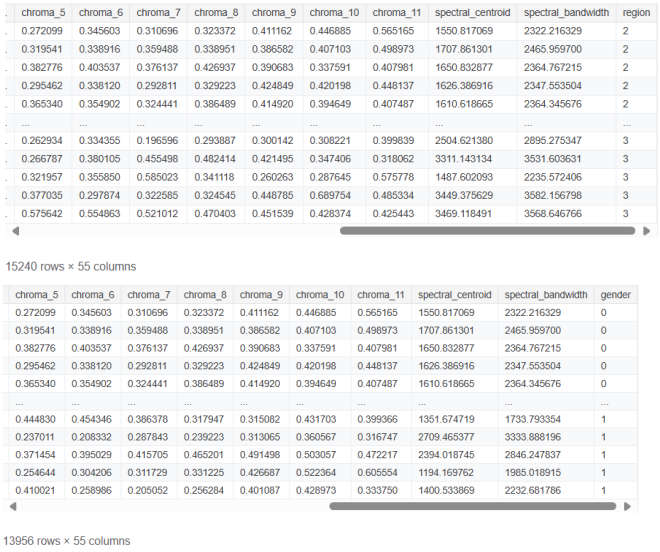


Fig. 4: Data Augmented and Balanced

The extracted features are converted to a 2D format (7x10x1) using the `make2dData` function, making them compatible with the CNN2D architecture.

3) Make 2-dimensional data.: **Preparing data for CNN2D** CNN2D requires 2D matrix data, so we have to do an additional step to convert 1D data to 2D according to the following steps: **Data normalization:**

- Use StandardScaler to normalize MFCC and Spectral features.
- Use MinMaxScaler to normalize Chroma features.
- Just fit the normalizers on the training set and then apply them to the dev and test sets.



	chroma_5	chroma_6	chroma_7	chroma_8	chroma_9	chroma_10	chroma_11	spectral_centroid	spectral_bandwidth	region
.	0.272099	0.345603	0.310696	0.323372	0.411162	0.446885	0.565165	1550.817069	2322.216329	2
.	0.319541	0.338916	0.359488	0.338951	0.386582	0.407103	0.498973	1707.861301	2465.959700	2
.	0.382776	0.403537	0.376137	0.426937	0.390683	0.337591	0.407981	1650.832877	2364.767215	2
.	0.295462	0.338120	0.292811	0.329223	0.424849	0.420198	0.448137	1626.386916	2347.553504	2
.	0.365340	0.354902	0.324441	0.386489	0.414920	0.394649	0.407487	1610.618665	2364.345676	2
.	...	...	...	...	...	...	...	...	...	...
.	0.262934	0.334355	0.196596	0.293887	0.300142	0.308221	0.399839	2504.621380	2895.275347	3
.	0.266787	0.380105	0.455498	0.482414	0.421495	0.347406	0.318062	3311.143134	3531.603631	3
.	0.321957	0.355850	0.585023	0.341118	0.260263	0.287645	0.575778	1487.602093	2235.572406	3
.	0.377035	0.297874	0.322585	0.324545	0.448785	0.689754	0.485334	3449.375629	3582.156798	3
.	0.575642	0.554863	0.521012	0.470403	0.451539	0.428374	0.425443	3469.118491	3568.646766	3

15240 rows × 55 columns

	chroma_5	chroma_6	chroma_7	chroma_8	chroma_9	chroma_10	chroma_11	spectral_centroid	spectral_bandwidth	gender
	0.272099	0.345603	0.310696	0.323372	0.411162	0.446885	0.565165	1550.817069	2322.216329	0
	0.319541	0.338916	0.359488	0.338951	0.386582	0.407103	0.498973	1707.861301	2465.959700	0
	0.382776	0.403537	0.376137	0.426937	0.390683	0.337591	0.407981	1650.832877	2364.767215	0
	0.295462	0.338120	0.292811	0.329223	0.424849	0.420198	0.448137	1626.386916	2347.553504	0
	0.365340	0.354902	0.324441	0.386489	0.414920	0.394649	0.407487	1610.618665	2364.345676	0
	...	...	...	...	...	...	...	...	...	...
	0.444830	0.454346	0.386378	0.317947	0.315082	0.431703	0.399366	1351.674719	1733.793354	1
	0.237011	0.208332	0.287843	0.239223	0.313065	0.360567	0.316747	2709.465377	3333.888196	1
	0.371454	0.395029	0.415705	0.465201	0.491498	0.503057	0.472217	2394.018745	2846.247837	1
	0.254644	0.304206	0.311729	0.331225	0.426687	0.522364	0.605554	1194.169762	1985.018915	1
	0.410021	0.258986	0.205052	0.256284	0.401087	0.428973	0.333750	1400.533869	2232.681786	1

13956 rows × 55 columns

Fig. 5: SRR and SGR data

Bundle normalized features: Combine the normalized MFCC, Chroma and Spectral features into a single input matrix for each dataset (train, dev, test). **Convert data from 1D to 2D:**

- Use the convert_1dto2d and make2Ddata functions to convert data from 1D to 2D.
- This function converts MFCC, Chroma and Spectral features into a 2D matrix of size (7, 10, 1).

The method is illustrated in Fig. 6.

D. Model Architecture and Training

We use a 2D Convolutional Neural Network (CNN2D) for both SRR and SGR tasks, with architectures similar to the optimized CNN2D model used for SER. The SRR model has an output layer with 3 units (North, Central, South), and the SGR model has 2 units (Male, Female). The architecture includes:

- **Input Layer:** (7, 10, 1) for 2D feature matrices.
- **Conv2D Layers:** Two convolutional layers with ReLU activation, followed by max-pooling.
- **Flatten and Dense Layers:** Flattening followed by dense layers with dropout and batch normalization for regularization.
- **Output Layer:** Softmax activation for classification (3 units for SRR, 2 units for SGR).

The models are trained using the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric. Training is conducted over 200 epochs with a batch size of 64.

E. Hyperparameter Tuning

To optimize model performance and reduce computational complexity, we use Keras Tuner with the Hyperband algorithm. The hyperparameter ranges are shown in Table II.

After 60 trials, the best hyperparameters are applied, as shown in Table III.

Hyperparameter	CNN2DRegion	CNN2DGender
Input Filters	Conv2D-1: 32-128 Conv2D-2: 16-64	Conv2D-1: 32-128 Conv2D-2: 16-64
Kernel Size	Conv2D-1: 3, 5 Conv2D-2: 3, 5	Conv2D-1: 3, 5 Conv2D-2: 3, 5
L2 Regularization	$1e^{-4}$ to $1e^{-2}$	$1e^{-4}$ to $1e^{-2}$
Dropout Rate	0.2-0.5	0.2-0.5
Pooling Size	Pool1: 2-3	Pool1: 2-3
Pooling Strides	Pool1: 1-2	Pool1: 1-2
Dense Units	32-128 16-64	32-128 16-64
Learning Rate	$1e^{-4}$ to $1e^{-2}$	$1e^{-4}$ to $1e^{-2}$

TABLE II: Hyperparameter Ranges for CNN2D Models (Region and Gender)

Parameter	Region	Gender
filters_0	80	32
kernel_size_0	5	5
l2_conv_0	0.0031038838	0.0046196045
pool_size_0	2	3
include_conv1	True	True
dropout_conv	0.4	0.2
include_dense0	True	False
units_1	8	16
l2_dense_1	0.0006723416	0.0013046810
dropout_dense_1	0.2	0.0
learning_rate	0.0002245907	0.0001646897
filters_1	48	48
kernel_size_1	3	5
l2_conv_1	0.0066001769	0.0006623877
units_0	64	48
l2_dense_0	0.0010304511	0.0027553943
dropout_dense_0	0.1	0.1
tuner/epochs	10	10
tuner/initial_epoch	0	4
tuner/bracket	0	2
tuner/round	0	2
tuner/trial_id	-	0014

TABLE III: Best Hyperparameters for Region and Gender CNN2D Models

III. Results

A. Training Results

The CNN2D models for SRR and SGR are trained on the augmented VA3R dataset. The training results are shown in Figures 7 and 8, with performance metrics summarized in Table IV.

Metric	CNN2DRegion	CNN2DGender
Epoch Stopped	88	20
Test Accuracy	99%	99%
Test Loss	0.09	0.14

TABLE IV: Initial Results from CNN2DRegion and CNN2DGender Models

The CNN2DGender model converges faster (epoch 20) compared to CNN2DRegion (epoch 88), but both achieve a test accuracy of 99% with low losses (0.09 for SRR, 0.14 for SGR).

B. Hyperparameter Tuning Results

After applying the optimized hyperparameters, the models are retrained, and the results are shown in Figures 9 and 10. The comparison of model size and convergence speed is presented in Table V.

Model	CNN2DRegion	CNN2DGender
Total params (Before)	1,765,995 (6.74 MB)	1,765,995 (6.74 MB)
Total params (After)	**758,619 (2.89 MB)**	**280,040 (1.07 MB)**
Epoch Stopped (Before)	88	20
Epoch Stopped (After)	**31**	**35**

TABLE V: Compare the models before and after applying hyperparameters

Hyperparameter tuning significantly reduces model size (from 6.74 MB to 2.89 MB for SRR and 1.07 MB for SGR) and speeds up convergence for SRR (from 88 to 31 epochs). For SGR, the number of epochs increases slightly (from 20 to 35), possibly due to oversimplification requiring more training to reach optimal performance.

C. Evaluation and Error Analysis

The final evaluation is conducted using the entire dataset as the test set. The results are shown in Figure 11, with confusion matrices in Figures 12 and 13, and classification metrics in Table VI.

Both models achieve an average Precision, Recall, and F1-score of 99.66%, indicating excellent performance. The SRR model correctly classifies most samples (e.g., 5,065 for North, 5,056 for Central, 5,067 for South), with minimal errors (e.g., 11 misclassifications for North). The SGR model also performs well, with 6,950 correct predictions for males and 6,959 for females, and only 28 and 19 misclassifications, respectively.

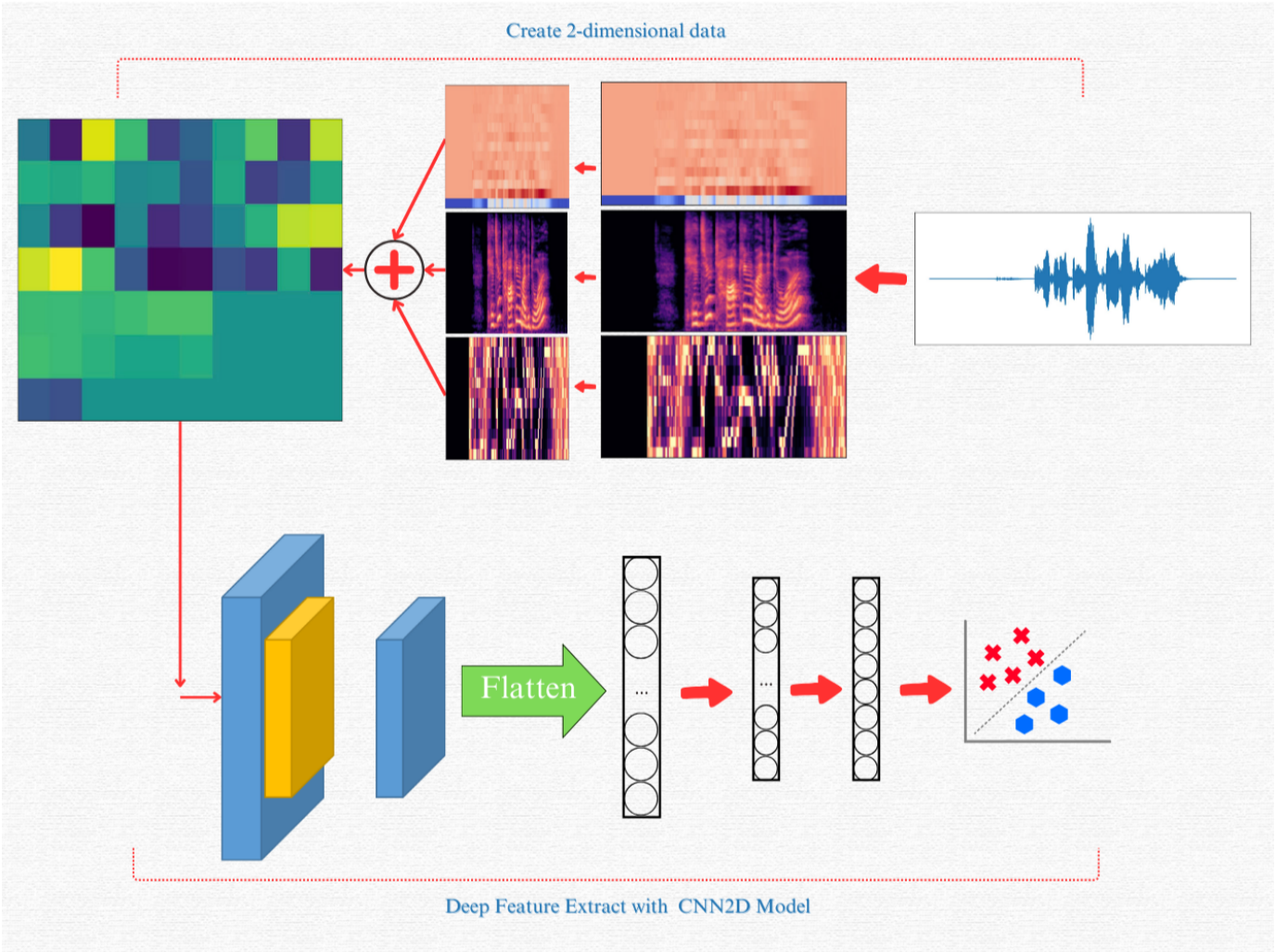


Fig. 6: Illustration of proposed method for CNN2D.

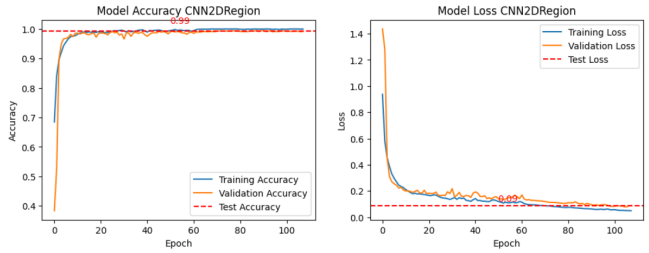


Fig. 7: CNN2DRegion model training results

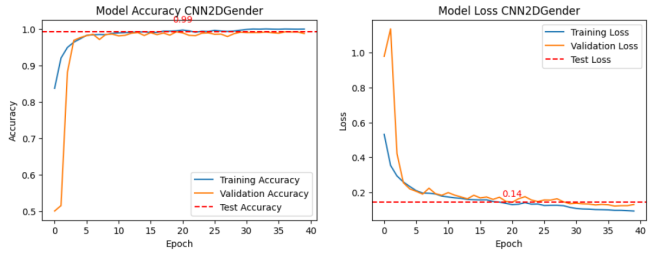


Fig. 8: CNN2DGender model training results

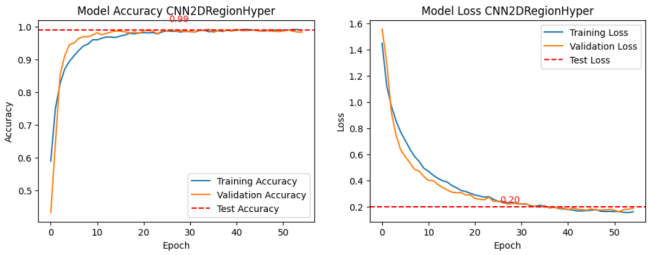


Fig. 9: CNN2DRegionHyper model training results

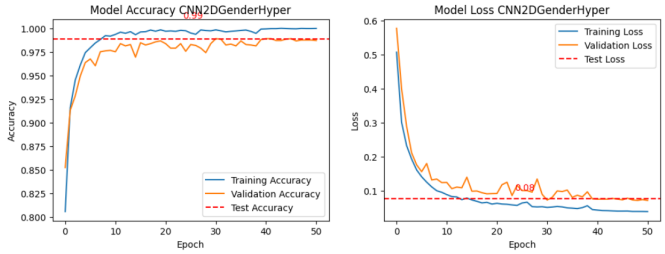


Fig. 10: CNN2DGenderHyper model training results

Class	Precision	Recall	F1-score
Region Classification			
0	0.9980	0.9970	0.9975
1	0.9963	0.9953	0.9958
2	1.0000	1.0000	1.0000
Average	**0.9966**	**0.9966**	**0.9966**
Gender Classification			
0	0.9973	0.9960	0.9966
1	0.9960	0.9973	0.9966
Average	**0.9966**	**0.9966**	**0.9966**

TABLE VI: Region and Gender model performance evaluation

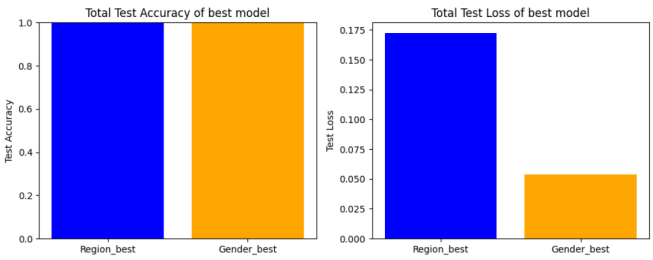


Fig. 11: Total Test Accuracy/Loss of best model

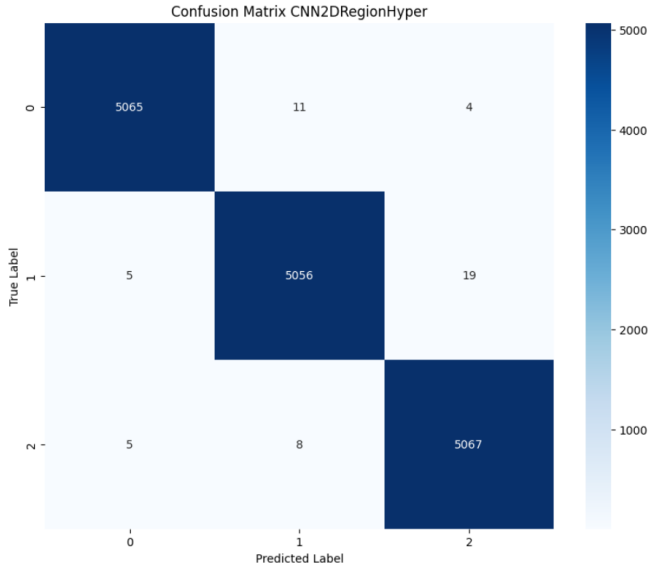


Fig. 12: Confusion matrix of CNN2DRegion model

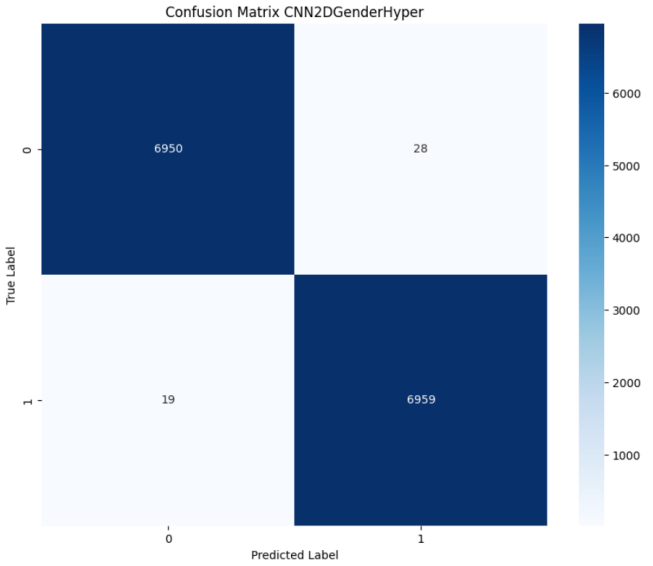


Fig. 13: Confusion matrix of CNN2DGender model

IV. Discussion

The proposed CNN2D models for SRR and SGR demonstrate exceptional performance, achieving 99.66% accuracy for both tasks. The use of loudness normalization ensures consistent input, while data augmentation addresses class imbalances, improving model robustness. Hyperparameter tuning with Keras Tuner significantly reduces model size and inference time, making the system suitable for real-time applications.

However, some limitations remain:

- **Dataset Diversity:** The VA3R dataset, while comprehensive, is limited to three regions and may not capture the full spectrum of Vietnamese accents (e.g., sub-regional variations).
- **Noise Robustness:** The models perform well on augmented data with controlled noise, but real-world scenarios with diverse background noise may pose challenges.
- **Model Complexity:** Although optimized, the CNN2D architecture may still be computationally intensive for low-resource devices.

Future work could explore:

- Expanding the VA3R dataset to include more regions and diverse noise conditions.
- Investigating lightweight architectures (e.g., MobileNet) for deployment on resource-constrained devices.
- Incorporating temporal features using Recurrent Neural Networks (RNNs) or Transformers to capture sequential patterns in speech.

V. System Integration and Demonstration

1. Integration of the Model into the Demonstration System

The system integrates three deep learning models for real-time speech classification, handling emotion, regional accent, and gender recognition. These models, trained as 2D CNNs, are stored in '.keras' format and loaded into a Flask-based application. The models are initialized and ready for inference upon system startup.

2. Real-Time Inference and Processing

Incoming audio is processed through a three-second sliding window, ensuring continuous real-time prediction. The microphone captures speech at a sampling rate of 48,000 Hz, with newly received audio appended to the buffer while older samples are removed. Every second, the system extracts features and updates predictions based on the latest three seconds of speech.

- **Audio Processing and Feature Extraction:** Before being processed by the models, the audio undergoes RMS normalization and feature extraction using Librosa. Extracted features include MFCCs (n=40), chromagram, spectral centroid, and spectral bandwidth. These features are converted into a CNN-compatible 2D format using task-specific transformation functions. Each classification task has a separate function, ensuring optimal preprocessing for emotion, accent, and gender recognition.
- **Sliding Window for Real-Time Predictions:** The system continuously updates the audio buffer, maintaining the last three seconds for inference. If fewer than three seconds of data are available, zero-padding is applied. Once the buffer is filled, extracted features are passed through the respective models to generate predictions, which are updated every second.
- **System Testing and Performance:** The system was tested on unseen speech samples to evaluate its accuracy and real-time performance. The emotion recognition model classifies speech into eight categories: Neutral, Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprised. The regional accent model identifies three dialects: Southern, Central, and Northern. The gender recognition model distinguishes between Male and Female speakers.
- **Inference Speed and Optimization:** The system processes each three-second speech window in approximately 0.33 milliseconds on a standard GPU, allowing near-instantaneous feedback. Optimization strategies such as parallel model execution and model quantization improve computational efficiency while maintaining accuracy.

3. Demo and Result Visualization

The system runs as a Flask web application hosted locally. The user interface displays real-time classification results, including detected emotions, accents, and gender, along with confidence scores and waveform intensity visualizations. The interface also includes sections for member information and user feedback.

The system interface includes a section introducing the team members. This section provides details about the contributors to the project and their roles.

Fig. 14: Members

The interface of the system displays the results of emotion, accent, and gender recognition in real time.

Fig. 15: system

The system also features a feedback section, allowing users to provide insights on its performance.

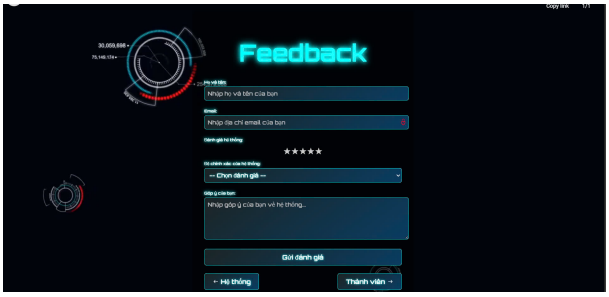


Fig. 16: Feedback

4. Testing the System with a Video Sample

A recorded video containing spontaneous speech was processed to validate system performance under real-world conditions. The extracted audio was segmented into three-second intervals and classified by the models. The system effectively recognized emotions, regional accents, and gender across different segments, demonstrating adaptability to natural speech variations. In the test,a YouTube video labeled as "Southern accent," the system detected it as "Southern" with 82.11% accuracy. The system correctly identified the speaker as male (99.67%) and classified the emotion as "calm" (99.99%). While gender and emotion recognition performed accurately, the regional accent classification may need further tuning. Possible factors include background noise or limitations in the training dataset. The system is functioning well in real-time but could be improved for better regional accent recognition.

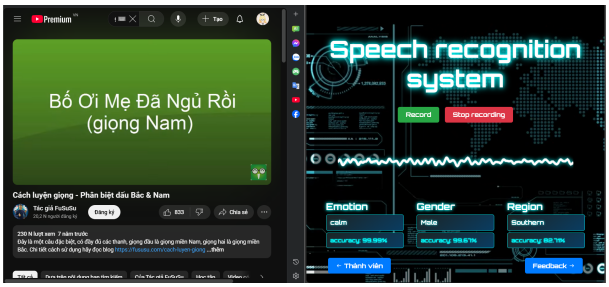


Fig. 17: Test_system

5. Conclusion and Future Work

This system enables real-time speech classification by integrating emotion recognition, regional accent detection, and gender identification. Future improvements will focus on enhancing accuracy, expanding training datasets, and further optimizing inference speed for broader deployment.

VI. Conclusion

This study successfully develops a deep learning-based system for Speech Region Recognition and Speech Gender Recognition using the VA3R dataset. The CNN2D models achieve high accuracy (99.66% for both SRR and SGR) through effective preprocessing, feature extraction, data augmentation, and hyperparameter optimization. The optimized models are lightweight and efficient, making them suitable for real-time applications such as speaker profiling and forensic analysis. Future improvements will focus on enhancing dataset diversity, noise robustness, and computational efficiency to broaden the system's applicability.

References

[1] yt-dlp, "yt-dlp: A youtube-dl fork with additional features and fixes," [Online]. Available: https://github.com/yt-dlp/yt-dlp

[2] Bredin, H., "pyannote-audio: Neural building blocks for speaker diarization and voice activity detection," [Online]. Available: https://huggingface.co/pyannote/voice-activity-detection

[3] Le Nhat Truong et al., "Vietnamese Accent 3 Regions (VA3R) Dataset," Kaggle, 2025. [Online]. Available: https://www.kaggle.com/datasets/lnt0821/vietnamese-accent-3-regions

[4] B. McFee, C. Raffel, D. Liang, D. P. W. 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 (SciPy)*, 2015.

[5] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[6] S. I. Levitan, T. Mishra, and S. Bangalore, "Automatic Identification of Gender from Speech," in *Proceedings of the Speech Prosody Conference*, 2016.

[7] H. T. Nguyen, Q. C. Nguyen, and T. P. Nguyen, "Vietnamese Dialect Identification Using Deep Neural Networks," in *Proceedings of the 2019 International Conference on Asian Language Processing (IALP)*, 2019.

[8] F. Chollet, "Keras: Deep Learning for Humans," [Online]. Available: https://keras.io

[9] A. G. Baydin, R. Cornish, D. M. Rubio, M. Schmidt, and F. Wood, "Hyperparameter Optimization with Keras Tuner," in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[10] Y. Li, M. Liu, and W. Zhang, "Speech Accent Recognition Using Deep Convolutional Neural Networks," in *IEEE International Conference on Signal Processing (ICSP)*, 2018.