Binary Speech Emotion Recognition: A Comparative Study of SVM, CNN, and Bi-LSTM Approaches

Project Proposal

Ammar Qurthuby*, Habibi*
*Department of Informatics Engineering, Universitas Syiah Kuala
Email: {ammar22, habibi123}@mhs.usk.ac.id

***Abstract*—This proposal presents a research plan for developing and comparing three machine learning approaches for binary speech emotion recognition (SER): Support Vector Machine (SVM), 1D Convolutional Neural Network (CNN), and Bidirectional Long Short-Term Memory (Bi-LSTM). We aim to classify emotions into two categories: Negative (angry, disgust, fear, sad) and Non-Negative (happy, neutral, surprise). Using a merged dataset combining CREMA-D, RAVDESS, SAVEE, and TESS containing approximately 12,000 audio samples, we will extract dual feature sets: statistical features for SVM and sequential MFCC features for deep learning models. The project will systematically compare traditional machine learning and deep learning approaches to identify the most effective method for binary emotion classification, with applications in mental health monitoring and human-computer interaction.**

***Index Terms*—Speech Emotion Recognition, Binary Classification, Support Vector Machine, Convolutional Neural Network, Bidirectional LSTM, MFCC Features, Mental Health Applications**

I. Introduction

Speech emotion recognition (SER) has become increasingly important in various applications including human-computer interaction, customer service systems, mental health monitoring, and intelligent personal assistants [1], [2]. The ability to automatically detect emotional states from speech signals can significantly enhance user experience and enable early detection of mental health conditions such as depression and anxiety [3].

Traditional SER systems typically classify emotions into multiple discrete categories (e.g., happy, sad, angry, neutral) based on dimensional models or categorical approaches [4]. However, for clinical and practical applications, binary classification of emotions into Negative and Non-Negative categories offers several advantages: (1) simplified decision-making for automated systems, (2) higher classification accuracy due to reduced complexity, and (3) direct clinical relevance for mental health screening [5].

Various machine learning approaches have been proposed for SER, ranging from traditional methods such as Support Vector Machines (SVM) and Random Forests to deep learning architectures including Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) [6]. While deep learning models have shown promising results on multi-class emotion recognition tasks, their performance on binary classification and the trade-offs between model complexity and accuracy remain relatively unexplored.

A. Research Objectives

This project aims to achieve the following objectives:

- Develop and implement three representative machine learning approaches for binary SER: SVM with RBF kernel, 1D-CNN, and Bidirectional LSTM
- Design and extract appropriate feature sets tailored to each model architecture
- Conduct systematic comparison of model performance using accuracy, precision, recall, and F1-score metrics
- Analyze the trade-offs between model complexity, training time, and classification performance
- Provide practical recommendations for deploying binary SER systems in real-world applications

B. Contributions

The expected contributions of this research include:

- A comprehensive comparative study of traditional machine learning (SVM) and deep learning approaches (CNN, Bi-LSTM) specifically for binary SER
- Analysis of dual feature extraction strategies: statistical features for traditional classifiers and sequential features for deep learning models
- Empirical insights into the effectiveness of different architectures for binary emotion classification
- Detailed performance analysis and practical deployment guidelines

II. Related Work

A. Speech Emotion Recognition Approaches

Speech emotion recognition has been extensively studied using various machine learning approaches. Traditional methods include Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), and Support Vector Machines [7]. El Ayadi et al. [2] provided a comprehensive survey showing that SVM with appropriate kernel functions consistently achieves competitive performance across different emotion recognition tasks.

Deep learning has gained prominence in SER in recent years. Convolutional Neural Networks have been successfully applied to spectrogram representations of speech [8], [9]. Recurrent architectures, particularly LSTM and Bi-LSTM, have shown promise in capturing temporal dependencies in speech signals [10], [11]. However, comparative studies show

mixed results regarding whether CNN or RNN architectures perform better for emotion recognition [6].

B. Feature Extraction for SER

Acoustic features play a crucial role in emotion recognition. Mel-Frequency Cepstral Coefficients (MFCC) remain the most widely used features due to their effectiveness in capturing spectral envelope information [12]. Other commonly used features include pitch, energy, zero-crossing rate, and spectral features [13].

Recent work has explored using raw waveforms with deep learning [14], but hand-crafted features like MFCC still provide strong baselines and require less computational resources. The choice between statistical aggregation versus sequential representation of features depends on the classification algorithm [15].

C. Binary vs. Multi-Class Classification

While most SER research focuses on multi-class classification (typically 4-8 emotion categories), binary classification has received less attention. Valstar et al. [5] highlighted the importance of binary valence (positive/negative) classification for clinical applications. Studies in depression detection have shown that binary emotion classification can achieve higher accuracy than fine-grained emotion recognition [3].

III. Proposed Methodology

A. Dataset

We will use a merged dataset combining four publicly available speech emotion databases: CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset), RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), SAVEE (Surrey Audio-Visual Expressed Emotion), and TESS (Toronto Emotional Speech Set) [16], [17], [18], [19].

The combined dataset will contain approximately 12,000 audio samples from diverse speakers with various demographic backgrounds. For binary classification, emotions will be grouped as:

- **Negative**: angry, disgust, fear, sad
- **Non-Negative**: happy, neutral, surprise

The dataset will be split into training (80%) and test (20%) sets using stratified sampling to maintain class balance.

B. Feature Extraction Strategy

We propose two different feature extraction approaches:

1) Statistical Features for SVM: For traditional machine learning approaches, we will extract 80-dimensional statistical feature vectors from each audio sample:

- **MFCC Statistics (40D):** Mean and standard deviation of 20 Mel-Frequency Cepstral Coefficients, capturing spectral envelope characteristics
- **Acoustic Features (40D):** Temporal statistics (mean, std) of zero-crossing rate, spectral centroid, spectral bandwidth, spectral rolloff, and RMS energy, providing complementary acoustic information

2) Sequential Features for Deep Learning: For CNN and Bi-LSTM models, we will extract sequential MFCC representations preserving temporal dynamics:

- **Feature Dimension:** 40 MFCC coefficients extracted per time frame
- **Sequence Length:** Fixed length of 100 time steps (padded or truncated as needed)
- **Representation:** Resulting 100×40 matrices per audio sample for temporal pattern learning

C. Proposed Model Architectures

1) Support Vector Machine:

- **Kernel:** Radial Basis Function (RBF) for non-linear classification
- **Hyperparameter Optimization:** Grid search with 5-fold cross-validation
- **Search Parameters:**
  - Regularization parameter C: Controls margin softness
  - Kernel coefficient γ: Controls decision boundary curvature
- **Class Weights:** Balanced weighting to handle potential class imbalance

2) 1D Convolutional Neural Network: The proposed CNN architecture will include:

- **Architecture:** Three convolutional blocks with progressively increasing filter depths (64, 128, 256 filters)
- **Regularization:** Batch normalization after each convolutional layer and max pooling for spatial dimensionality reduction
- **Dimensionality Reduction:** Global average pooling to minimize parameters and prevent overfitting
- **Classification Layers:** Fully connected dense layers with dropout (rate: 0.5) for final classification
- **Optimization:** Adam optimizer with initial learning rate of 0.001 and early stopping mechanism

3) Bidirectional LSTM: The proposed Bi-LSTM architecture will consist of:

- **Architecture:** Three stacked bidirectional LSTM layers with decreasing units (128, 64, 32 units per direction)
- **Temporal Processing:** Bidirectional processing to capture both past and future context in speech sequences
- **Regularization:** Dropout (rate: 0.3) between LSTM layers to prevent overfitting
- **Classification Layers:** Fully connected dense layers for mapping LSTM outputs to emotion classes
- **Optimization:** Adam optimizer with lower learning rate of 0.0001 for stable convergence

D. Evaluation Plan

Models will be evaluated using:

- **Performance Metrics**: Accuracy, Precision, Recall, F1-Score
- **Confusion Matrix**: Detailed error analysis

- **Training Efficiency**: Training time and convergence behavior
- **Statistical Testing**: Results averaged over 3 independent runs

IV. Expected Outcomes and Timeline

A. Expected Results

Based on literature review, we anticipate:

- All three models achieving ¿75% accuracy on binary classification
- Deep learning models potentially outperforming SVM due to automatic feature learning
- Trade-offs between model complexity and computational efficiency
- Insights into which architecture is most suitable for binary SER

B. Project Timeline

TABLE I
Project Timeline

Week	Activities
1-2	Dataset collection and preprocessing
3-4	Feature extraction implementation
5-6	SVM model development and tuning
7-8	CNN model development and training
9-10	Bi-LSTM model development and training
11-12	Results analysis and comparison
13-14	Documentation and final report

C. Resources Required

- **Hardware**: Computer with GPU support (NVIDIA GTX 1660 Ti or equivalent)
- **Software**: Python 3.8+, TensorFlow/Keras, scikit-learn, librosa
- **Datasets**: CREMA-D, RAVDESS, SAVEE, TESS (all publicly available)
- **Computing Time**: Estimated 40-50 hours for training all models

V. Conclusion

This proposal outlines a comprehensive research plan for comparing three machine learning approaches for binary speech emotion recognition. By systematically evaluating SVM, 1D-CNN, and Bi-LSTM on a large merged dataset, we aim to provide practical insights for developing emotion recognition systems. The dual feature extraction strategy allows fair comparison between traditional and deep learning methods. Expected outcomes include identifying the most effective approach for binary SER and providing deployment guidelines for real-world applications in mental health monitoring and human-computer interaction.

Acknowledgment

We would like to thank the Department of Informatics Engineering, Universitas Syiah Kuala, for providing the computational resources and guidance for this research project.

References

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

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

3 Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F., "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

4 Ekman, P., "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

5 Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

6 Zhang, S., Zhang, S., Huang, T., and Gao, W., "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2018.

7 Nwe, T. L., Foo, S. W., and De Silva, L. C., "Speech emotion recognition using hidden markov models," *Speech communication*, vol. 41, no. 4, pp. 603–623, 2003.

8 Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W., "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 International Conference on Platform Technology and Service (PlatCon)*. IEEE, 2017, pp. 1–5.

9 Sarma, M., Ghahremani, P., Povey, D., Goel, N. K., Sarma, K. K., and Dehak, N., "Emotion identification from raw speech signals using DNNs," in *Proc. Interspeech 2018*, 2018, pp. 3097–3101.

10 Mirsamadi, S., Barsoum, E., and Zhang, C., "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

11 Huang, Z., Dong, M., Mao, Q., and Zhan, Y., "Speech emotion recognition using deep neural network and extreme learning machine," *Frontiers in neurorobotics*, vol. 8, p. 34, 2014.

12 Davis, S. and Mermelstein, P., "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.

13 Schuller, B., Steidl, S., and Batliner, A., "Acoustic emotion recognition: A benchmark comparison of performances," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE, 2009, pp. 552–557.

14 Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

15 Eyben, F., Wöllmer, M., and Schuller, B., "openSMILE: the Munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.

16 Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R., "CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

17 Livingstone, S. R. and Russo, F. A., "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.

18 Haq, S. and Jackson, P. J., "Speaker-dependent audio-visual emotion recognition," in *Proceedings of the 4th International Conference on Auditory-Visual Speech Processing (AVSP)*. Tangalooma, Australia, 2008, pp. 53–58.

19 Dupuis, K. and Pichora-Fuller, M. K., "Toronto Emotional Speech Set (TESS)," Scholars Portal Dataverse, 2010, available at: https://tspace.library.utoronto.ca/handle/1807/24487.