Text and Voice Integration with Wav2Vec and Specaugment for Improving Accuracy of Speech Emotion Recognition

1st Edward Soenardi
Informatics Department
Petra Christian University
Surabaya, Indonesia
c14210097@john.petra.ac.id

2nd Djoni Haryadi Setiabudi
Informatics Department
Petra Christian University
Surabaya, Indonesia
djonihs@petra.ac.id

3rd Indar Sugiarto
Department of Electrical Engineering
Petra Christian University
Surabaya, Indonesia
indi@petra.ac.id

2025 International Conference on Data Science and Its Applications (ICoDSA) | 979-8-3315-9854-9/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICoDSA67155.2025.11156967

***Abstract*—Emotions are crucial in understanding the human psychological state and can influence social interactions and decision-making. The integration between emotion processing and computational systems enables the creation of more natural and adaptive interfaces. This research focuses on Speech Emotion Recognition (SER), which aims to identify human emotions through voice signal analysis. Many previous studies have been limited to unimodal approaches, and just few have explored multimodal approaches that combine voice and text simultaneously. Furthermore, there is a lack of research that directly integrates Automatic Speech Recognition (ASR) for text as well as voice feature extraction in one step and applies data augmentation to improve model generalization. This research contributes to the improvement of accuracy for emotion recognition tasks using the IEMOCAP dataset with total 5797 voices and text that contain angry, sad, happy, and neutral by incorporating the Wav2Vec model as a multimodal feature extractor (voice and text), and by applying SpecAugment to enrich the data variety. Structurally, our proposed architecture consists of two branches: a voice branch and a text branch. Features extracted from Wav2Vec are sent to the voice branch using an ECAPA-TDNN model, and to the text branch using a BERT model. These two branches are then combined using fully connected layers for final classification. Experiments show that this multimodal approach can achieve high performance, namely weighted accuracy of 90.28% and unweighted accuracy of 90.62%, with requiring special fine-tuning of the text model. These results indicate that the integration of multimodal with pretrained and data augmentation approaches can significantly improve the performance of SER systems.**

Keywords—voice signal analysis, speech emotion recognition, multimodal, feature extraction, Wav2Vec

I. Introduction

Emotions play a crucial role in human life, influencing cognitive processes, behavior, and social interactions [1]. The ability to perceive and respond to the emotions of others enhances the quality of communication [2] and supports better decision making [3]. In the context of human-computer interaction, recognizing emotions can facilitate more natural and effective communication, thereby improving user experience in a wide range of applications [4].

In order to understand and analyze these emotions through voice, the development of Speech Emotion Recognition (SER) systems is becoming increasingly relevant, especially in supporting various applications, such as virtual assistants, automated customer service, and AI-based health technology [5]. However, one of the core challenges in SER is that human emotions are often expressed through multiple modalities simultaneously, including vocal characteristics, linguistic content, facial expressions, and body gestures [6]. This complexity makes it difficult to achieve high accuracy using a single modality alone. Previous studies have shown that unimodal approaches particularly those using only audio typically achieve weighted and unweighted accuracy between 70%-77% [7], [8], [9], [10]. This limitation arises from the difficulty in distinguishing between emotions such as anger, happiness, and sadness based solely on vocal tone, and the challenge of identifying emotions like sadness using only textual content [11].

To address these issues, this research proposes a multimodal approach to SER by combining both speech and text modalities. Traditionally, features have been extracted from audio using methods such as Mel-spectrogram [12], Mel-frequency cepstral coefficients (MFCCs)[13], Glove [14], and Hubert [15]. These approaches often required separate Automatic Speech Recognition (ASR) systems to transcribe audio into text for linguistic analysis [15]. Therefore, a method is needed to convert voice into text and retrieve voice features at one time to generate emotion predictions [16], and the next problem is that there are very few labeled speech emotion recognition datasets that can affect the model training process [5], [17] .

The primary contribution of this work is to enhance emotion recognition performance, specifically for the emotions of angry, sad, happy, and neutral, through an integrated multimodal architecture. Previous multimodal study predicting these four emotions has achieved weighted accuracy and unweighted accuracy of 78% using ResTDNN for voice model and BERT for text [14].

This research aims to evaluate the performance by combining two branch models containing BERT model for analyzing text and ECAPA-TDNN model for analyzing speech data and investigate the performance improvement of the dataset with the addition of augmentation data using SpecAugment. The two branches' data will be obtained from Wav2Vev to extract speech features and to transcribe text at the same time. To address the gap, this research applies augmentation with SpecAugment to increase the variety of data [18] and uses Wav2Vec as a model to extract more representative speech and text features [19]. The dataset to be used is IEMOCAP, which specifically contains speech emotion data.

The idea in this research begins with training and evaluating a speech model using ECAPA-TDNN, applied separately to four types of datasets: IEMOCAP, IEMOCAP with augmentation, RAVDESS, and RAVDESS with augmentation. This is followed by training and evaluating a text model using BERT on two types of datasets: IEMOCAP and a combination of IEMOCAP and ISEAR. Finally, the integration of both modalities—speech and text—is performed using two types of datasets that include IEMOCAP and its augmented versions. For each dataset, four types of integrated models are trained and evaluated: (1) without any pretraining, (2) using text pretrained weights, (3) using speech pretrained weights, and (4) using both speech and text pretrained weights. The pretrained weights are selected based on the best-performing single-modality models.

This research not only demonstrates improved accuracy over previous multimodal approaches, but also shows that Wav2Vec-based multimodal representation combined with data augmentation significantly enhances emotion recognition, providing a more unified and efficient pipeline for real-world SER applications.

II. Related Work

Several previous studies related to speech emotion recognition with singe-modality have been conducted using IEMOCAP dataset. For example, E. Morais et al. [10] conducted research using voice only with Wav2vec/Hubert as the feature extraction model to detect four emotions (happy, sad, angry, neutral) using so-called Emphasized Channel Attention, Propagation and Aggregation in Time Delay Neural Network (ECAPA-TDNN). The results obtained unweighted accuracy of 77.76% and weighted accuracy of 77.36%. These results show that wav2vec can do feature extraction quite well and the ECAPA-TDNN accuracy is quite good as well. J. Wang et al. [8] conducted research with single modality (voice only) using MFCCs and Mel-spectrograms as feature and using Dual LSTM model. Their research provides unweighted accuracy of 73.30% and a weighted accuracy of 72.70%. The shortcoming in their study is that audio signals are converted into MFCC and Mel-spectrograms is not enough to achieve the highest potential due to loss of emotional information. Z. Huijuan et al. [20] conducted research of voice only using 3D log-Mel as feature and CNN blocks with RNN attention module for the model. Their research provides the best model 3D-HTML with unweighted accuracy of 46% and weighted accuracy of 47%. Z. Yao et al. [12] conducted research of voice only using Mel-spectrogram as the input and fused three models deep learning contains HSF-DNN, MS-CNN and LLD-RNN. Their research provides unweighted accuracy of 58% and a weighted accuracy of 57%. Md Shah et al. [21] conducted voice only research using MFCCs and epoch-based features that yield unweighted accuracies of 64.2%. Mingke Xu et al. [22] conducted voice only research also using MFCCs for input and ACNN model that yield unweighted accuracies and weighted accuracies of 76% each.

Several previous studies related to speech emotion recognition with multi-modality have been conducted using IEMOCAP. For example, G. Sahu [11] conducted research voice only, text only and voice and text using random forest, XGBoost, SVM, Multinomial Naïve Bayes, Logistic Regression, Multi Layer Perceptron dan LSTM to detect four emotions (happy, sad, angry, and neutral). The results accuracy are in a range from 56 to 70%, which are not much different from accuracy obtained using other deep learning methods. Using only voice, it is very difficult to detect happy, sad and angry emotions. If using only text, it is difficult to detect sad emotions. W. Wu et al. [14] conducted research using both voice and text and using Glove as well as transcription with no ASR and Residual Time Delay Neural Network (ResTDNN) as voice model and BERT as text model. Two models will then be combined using a fully connected network. The results obtained are the best with an unweight accuracy of 78.41% and weighted accuracy of 77.57%. However, the drawback in their research was that it extracts text and voice features separately. Y. Wang et al. [23] conducted similar research but using Mel-Frequency Ceptral Coefficients (MFCC), chroma, pitch, zero-crossing rate, spectral and statistical measures as the feature and using Multimodal Transformer augmented fusion model. The results obtained unweighted accuracy of 72% and weighted 75%. Seunghyun Yoon et al. [24] conducted research on voice and text using dual recurrent neural networks (RNNs) and then combined the information from these sources to predict the emotion classes. The results accuracies obtained ranging from 68.8% to 71.8%.

III. Methods

We propose a method that require five steps in the development of the model for integrating text and speech for predicting emotion through speech. Those steps, as shown in Fig. 1, are data collection, augmenting and preprocessing, modeling, training, and evaluating.

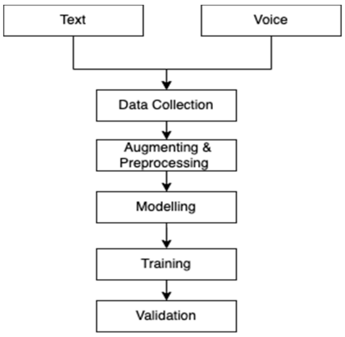


Fig. 1. Steps in the development of the model for integrating text and speech.

A. Data Collection

This is the first step in this research to obtain the datasets that will be used in model training and evaluation using English-language data only. This study uses two main datasets, namely IEMOCAP and RAVDESS for the speech, which are collected from sail.usc.edu [25] and kaggle [26], and ISEAR for additional dataset for training text model which are collected from kaggle [26].

B. Augmenting and Preprocessing

Before the dataset is used for training, it is necessary to augment & preprocess the dataset to ensure the data is more varied, cleaner, consistent, and ready to be processed. The augmentation stage is the process of adding new variations to existing data artificially to increase the robustness of the model. The augmentation dataset will be used SpecAugment, which is applied only to the audio data.

The preprocessing stage is essential to ensure data consistency and prevent issues such as null values that may cause errors during model training. This stage begins with filtering the RAVDESS and IEMOCAP datasets, which originally have seven emotion categories, down to four key emotions: angry, sad, happy, and neutral, based on the scope of this research. No speaker overlaps over RAVDESS and

IEMOCAP. For the IEMOCAP dataset specifically, the "excited" label is grouped under the "happy" category. Wav2Vec is then used for feature extraction—converting audio into latent features for the speech model, while also transcribing the audio into text for the text-based model.

Especially for the transcribe results of this extracting will be processed again. The text preprocessing stage will involve cleaning the text from HTML tags, digits, underscores, and stop words. Removal of links, special characters, non-ASCII characters, emails, and punctuation. As well as converting all text to lowercase to equalize formatting.

C. Modelling

There are three types of models that will be used for training, which are text model, voice model, and text and voice model.

a) Voice Model: The process begins with the extraction of voice feature vectors from wav2vec, which are directly processed by the ECAPA-TDNN model. The model analyzes the acoustic features to generate an emotion prediction output based on the audio input.

b) Text Model: In Fig. 2, the process begins with the transcription input, represented as input IDs and attention masks in matrix form. These inputs are passed to the BERT model, where token-level analysis of the text is performed. The resulting hidden states from the BERT model are then dimensionally transformed to enable processing through three one-dimensional convolutional layers, each with kernel sizes of 3, 4, and 5. After the convolution operations, a pooling operation is applied, and the resulting feature maps are concatenated. This concatenated output is then passed through two fully connected layers, followed by a dropout layer for regularization. The text model used in this architecture was already fine-tuned using the ISEAR dataset specifically for emotion recognition.

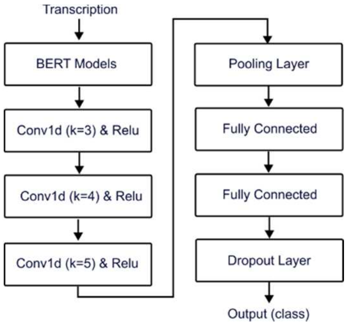


Fig. 2. Model text architecture.

c) Text and Voice Model: In Fig. 3 the process begins by extracting textual data from speech using Wav2Vec, which is then fed into the BERT model to predict emotions based on textual content. Simultaneously, acoustic features are also extracted from the same Wav2Vec output and passed to the ECAPA-TDNN model to predict emotions from the audio signal. Once both voice and text have been analyzed, the next step is the integration stage, where the embeddings from the two modalities are combined and passed through a dropout layer. The resulting representation is then input into a fully connected layer, which produces the final emotion predictions in the form of logits. This approaches its called single-pass Wav2Vec for dual features [27]. It uses a fully connected layer concatenation of BERT and ECAPA-TDNN embeddings before the final classifier , design choice made to reduce complexity and training time while still allowing the model to learn joint representations effectively.

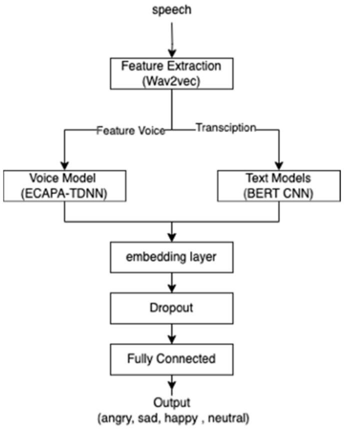


Fig. 3. Integrating voice and text architecture.

D. Training

The dataset was initially divided into two subsets: 80% for training and 20% for testing. To ensure a fair evaluation and prevent data leakage, we confirmed that there was no speaker overlap between the training and testing sets. Since this study conducts experiments separately on the RAVDESS and IEMOCAP datasets, each dataset was split independently. For both datasets, we ensured that speakers in the training set do not appear in the test set. The training process was conducted over 50 epochs using a batch size of 64. Each batch underwent several key stages. First, a forward pass was performed, where the model processed the input to generate predicted outputs. Subsequently, the CrossEntropy Loss function was used to compute the loss by measuring the difference between the predicted output and the ground truth labels. Backward propagation was then applied to calculate gradients, which were used to update the model's weights. Optimization was carried out using the Adam optimizer with a learning rate of 0.001, enabling the model to adjust its parameters for better pattern recognition. After each parameter update, the training accuracy was computed to evaluate the model's performance during the training phase. Following each epoch, the model was validated using the testing set to assess its ability to generalize to unseen data. An early stopping mechanism was employed to prevent overfitting, which halted training if no improvement in validation loss was observed. Specifically, early stopping was triggered if the validation loss did not improve by at least 0.01 over three consecutive epochs. The model with the best validation loss was saved as a .pth file to the designated path.

E. Validation

The validation process is conducted using the test data that was previously set aside during the dataset split. The predicted emotion labels are evaluated using both weighted accuracy and unweighted accuracy metrics, as the dataset contains imbalanced emotion classes. Additionally, a confusion matrix is used to compare the predicted labels against the true labels, providing further insight into the model’s performance in detecting emotions from speech.

IV. Experimental Results

In this section, we present the results of the training and testing processes for three models: the Voice model, the Text model, and the Integrated Voice and Text model, all trained using English-language data only from IEMOCAP .

TABLE I. TRAINING PERFORMANCE

	Dataset	Type	Epoch	Loss	WA	UA
a	RAV	AO	14	0,02	99.63%	99.68%
b	RAV + AU	AO	9	0.03	99.35%	99.35%
c	IEM	AO	8	0,87	71.74%	72.30%
d	IEM + AU	AO	7	0,98	63.71%	64.24%
e	IEM	TO	10	2,73	51.26%	13.89%
f	IEM + ISE	TO	11	2,70	54.25%	21.66%
g	IEM	AT+NP	7	0,32	89.31%	89.50%
h		AT+VP	5	0,77	79.75%	79.82%
i		AT+TP	6	0,33	89.66%	90.07%
j		AT+VP+TP	4	1.18	73.97%	73.68%
k	IEM + AUG	AT+NP	5	0,33	89.31%	89.48%
l		AT+VP	8	0,46	88.08%	88.13%
m		AT+TP	8	0,24	92.13%	92.25%
n		AT+VP+TP	10	0,44	88.38%	88.42%

a. AO: audio-only,TO: text-only,AT: Audio + Tex, NP: no pretrained, VP: voice pretrained, TP: Text pretrained , VP+TP: voice and text pretrained, AU: augmentation, IEM: IEMOCAP, RAV: RAVDESS, ISE: ISEAR

A. Voice model

The training results for the voice-based emotion recognition models are summarized in Table I (rows a - d). The training process for each configuration was terminated early once the model achieved a training accuracy above 80%, before 14 epochs. The best training performance was observed in the RAVDESS dataset (row a), where the model achieved exceptionally high weighted accuracy (WA) and unweighted accuracy (UA) of approximately 99% after 14.

TABLE II. VALIDATION PERFORMANCE

	Dataset	Type	Loss	WA	UA
a	RAV	AO	0,95	77.04%	75.83%
b	RAV+AU	AO	0,45	88.48%	88.24%
c	IEM	AO	1.68	58.35%	58.21%
d	IEM+AU	AO	1,42	62.12%	63.05%
e	IEM	TO	2,03	69.48%	69.91%
f	IEM+ISE	TO	2,12	72.65%	71.91%
g	IEM	AT+NP	0,94	72.54%	72.97%
h		AT+VP	1,14	70.28%	70.92%
i		AT+TP	0,57	84.72%	84.99%
j		AT+VP+TP	0,62	84.72%	83.90%
k	IEM+AU	AT+NP	0,39	87.27%	87.96%
l		AT+VP	0,70	87.40%	87.84%
m		AT+TP	0,33	90.28%	90.62%
n		AT+VP+TP	0,47	90.24%	90.22%

b. AO: audio-only,TO: text-only,AT: Audio + Tex, NP: no pretrained, VP: voice pretrained, TP: Text pretrained , VP+TP: voice and text pretrained, AU: augmentation, IEM: IEMOCAP, RAV: RAVDESS, ISE: ISEAR

After the training process was completed, the trained model was validate using test data. The model validated using voice data achieves a validation accuracy between 59% - 88%, which can been seen in Table II rows a - d. The best performance in audio-only is RAVDESS dataset with augmentation (row b) because RAVDESS is easier to learn due to its clean and consistent data structure. Also with augmentation, it can learn more general.

In Table II, IEMOCAP presents greater challenges due to high variation in emotions, noise, and natural interactions between speakers. As a result, the model trained on IEMOCAP (row c) exhibited lower validation performance, with WA dropping to 58.35% and UA as low as 58.21%. However, after applying data augmentation (row d), validation UA improved to 63% and WA improved to 62%, highlighting the benefit of augmentation techniques in handling imbalanced and noisy data.

B. Text Model

The training results for the text-based emotion recognition models are summarized in Table 1 (rows e - f). The training process for each configuration was terminated early once the model achieved a training accuracy above 50%, before 11 epoch. The best training performance was observed in the IEMOCAP and ISEAR combination dataset (row e), where the model achieved exceptionally high weighted accuracy (WA) and unweighted accuracy (UA) of approximately 54% and 21% after 11.

After the training process was completed, the trained model was validate using test data. The model validate using text data achieved a validation accuracy above 72% which can been seen in Table II rows e - f. The best performance in text-only is IEMOCAP and ISEAR combination dataset (row e) indicating that the inclusion of the ISEAR dataset contributed to improved generalization.

C. Integrating Voice and Text Model

The training results for the integrating voice and text emotion recognition models are summarized in Table I (rows g - n). The training process for each configuration was terminated early once the model achieved a training accuracy above 70%, before 11 epoch. The best training performance was observed in the IEMOCAP with augmentation dataset (row m) using text pretrained, where the model achieved exceptionally high weighted accuracy (WA) and unweighted accuracy (UA) of approximately 90% and 90% at 8 epoch.

After the training was completed, the trained model was validated using test data. The model validation using voice and text data achieved a validation accuracy above 70% for no augmentation and above 88% for additional augmentation which can been seen in Table II rows g - n. The best performance in integrating voice and text is IEMOCAP with augmentation and using text pretrained (row m) indicating that augmentation makes data more divers and robust.

However, some misclassifications were observed, particularly between neutral, sad, and happy emotions (see Fig. 4 bottom left). For instance, in several cases, calm or low-energy happy speech (e.g., soft laughter or mild joy) was incorrectly classified as neutral due to its subtle vocal cues.

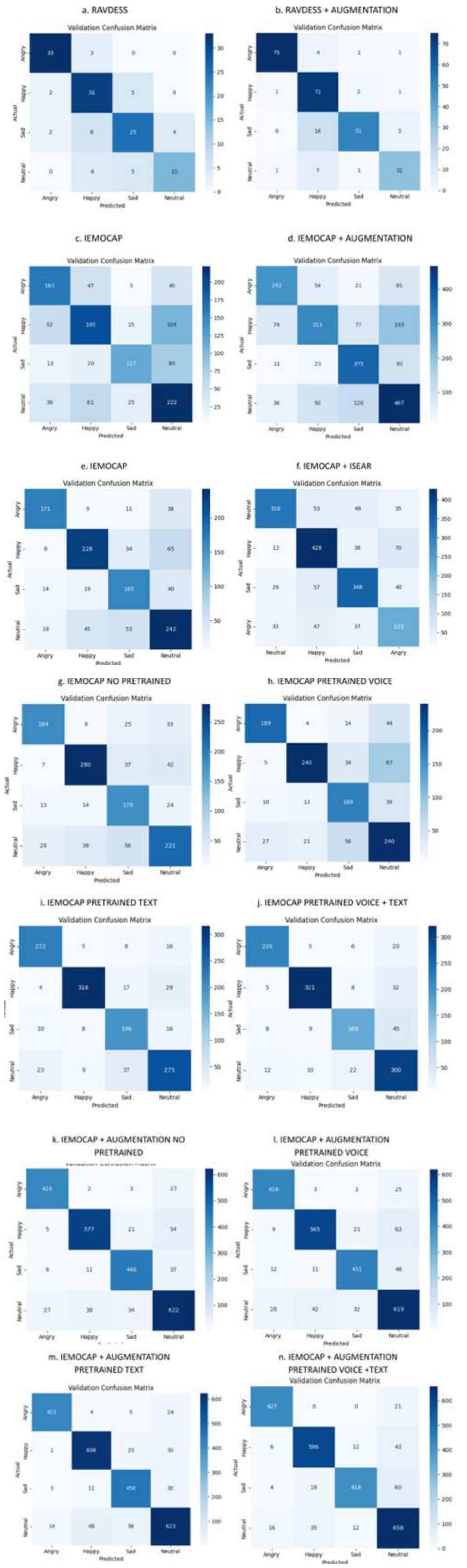
Similarly, soft-spoken sadness with minimal prosodic variation was occasionally labeled as neutral. In a few edge cases, emotionally ambiguous utterances such as a neutral sentence spoken with a slightly upbeat tone led the model to predict 'happy' instead of 'neutral.' These examples highlight the difficulty of distinguishing between low-intensity emotions and emphasize the need for richer contextual cues or more fine-grained emotion categories.

V. Discussion

Our results demonstrate that the proposed multimodal model achieves approximately an 8% improvement in emotion recognition accuracy compared to previous research [23] by leveraging augmented datasets and pretrained text models (row m). The model's ability to extract features using Wav2Vec captures both latent voice characteristics and simultaneous transcription before processing by ECAPA-TDNN and BERT, contributing to robust and varied data representation. This is reflected in strong per-emotion F1 scores, with Angry reaching 0.9421, Happy 0.9140, Sad 0.8968, and Neutral 0.8725, highlighting consistent performance across all classes. The model's training time was 423 minutes, with a size of 450 MB, and an average inference speed of 1 minute 3 seconds, demonstrating a practical balance between accuracy and computational efficiency suitable for real-time applications.
 However, using Wav2Vec as a feature extractor for single-modality input either text or audio led to signs of under 70% accuracy, as shown in Table I and Table II (rows a - d). This limited performance can be attributed to the inherent ambiguity in emotional cues when relying on only one modality. For example, vocal features alone may not fully capture the semantic intent behind emotionally neutral words, while textual input may miss the prosodic and tonal variations essential to interpreting emotions. In contrast, when the text and voice modalities were combined, overall model performance improved significantly in terms of both weighted and un-weighted accuracy.

As illustrated in Fig. 4, which presents the confusion matrix results, the audio-only model struggled to distinguish between similar emotions, particularly happy and sad with neutral in Fig. 4c and Fig. 4d, due to overlapping acoustic patterns. Meanwhile, the text-only model is often misclassified as happy and angry emotions in Fig. 4f. This limitation arises from the fact that text data cannot capture intonation, pitch, and vocal intensity, which are crucial for distinguishing such emotional nuances resulting in angry being frequently misclassified as happy.

When the model integrated both audio and text features, the combined strengths of each modality helped to reduce misclassification errors in Fig. 4m. The text provided semantic context, while the audio conveyed prosodic cues, leading to a more robust and accurate emotion recognition system. Deploying ASR-based Speech Emotion Recognition (SER) systems in real-world settings raises privacy concerns, as sensitive spoken data may be captured and processed, requiring strong data protection and anonymization measures. Despite this, the fusion of Wav2Vec and BERT achieves an average inference latency of just over one minute, demonstrating promising potential to meet real-time constraints with further optimization. Additionally, if used in another language, the results may differ and fall outside the scope of this study, which is based on English.



a. RAVDESS

b. RAVDESS + AUGMENTATION

c. IEMOCAP

d. IEMOCAP + AUGMENTATION

e. IEMOCAP

f. IEMOCAP + ISEAR

g. IEMOCAP NO PRETRAINED

h. IEMOCAP PRETRAINED VOICE

i. IEMOCAP PRETRAINED TEXT

j. IEMOCAP PRETRAINED VOICE + TEXT

k. IEMOCAP + AUGMENTATION NO PRETRAINED

l. IEMOCAP + AUGMENTATION PRETRAINED VOICE

m. IEMOCAP + AUGMENTATION PRETRAINED TEXT

n. IEMOCAP + AUGMENTATION PRETRAINED VOICE +TEXT

Validation Confusion Matrix
Actual
Predicted
Angry
Happy
Sad
Neutral
a: 33 3 0 0 / 2 31 5 0 / 2 6 25 4 / 0 4 5 15
b: 75 4 1 1 / 1 72 2 1 / 0 16 51 5 / 1 7 1 31
c: 161 47 3 40 / 52 195 15 104 / 13 20 117 90 / 36 61 25 222
d: 292 54 21 85 / 74 313 77 193 / 11 23 373 93 / 36 92 126 467
e: 171 9 11 38 / 8 228 34 65 / 14 19 165 40 / 18 45 53 242
f: 316 53 48 35 / 13 428 36 70 / 26 57 348 40 / 33 47 37 223
g: 184 9 25 33 / 7 280 37 42 / 13 14 179 24 / 28 39 56 221
h: 189 4 14 44 / 5 240 34 87 / 10 13 169 39 / 27 21 56 240
i: 222 5 8 16 / 4 316 17 29 / 10 8 196 16 / 23 9 37 275
j: 220 5 6 20 / 5 321 8 32 / 8 9 169 45 / 12 10 22 300
k: 416 2 3 27 / 5 577 21 54 / 6 11 446 37 / 27 38 34 622
l: 419 3 2 25 / 9 565 21 62 / 12 11 431 46 / 28 42 32 619
m: 411 4 5 24 / 1 606 20 30 / 3 11 456 30 / 14 48 36 623
n: 427 0 0 21 / 6 596 12 43 / 4 18 418 60 / 16 35 12 658

Fig. 4. Confusion matrix for voice only (a-d),text only (e-f) , and integrating voice and text (g-k).

VI. Conclusions

In this paper, we proposed a multimodal approach to Speech Emotion Recognition (SER) by integrating both voice and text modalities. Leveraging Wav2Vec for joint feature extraction, we demonstrated that using Wav2Vec as a common backbone for both audio and textual inputs can yield strong performance when combined with SpecAugment, which enhances data robustness and variability. Our experiments on the IEMOCAP dataset show that while Wav2Vec alone is suboptimal for unimodal emotion recognition, its effectiveness significantly improves when integrated with BERT for text and ECAPA-TDNN for audio. This study achieved state-of-the-art performance by combining text-pretrained models with data augmentation, achieving a weighted accuracy of 89.21% and an unweighted accuracy of 89.98%, demonstrating the effectiveness of our multimodal approach. These results highlight the importance of multimodal learning and data augmentation in improving the performance and generalizability of SER systems. Future work may explore the use of cross-modal transformers or fusion strategies to further enhance emotion recognition across diverse datasets and real-world scenarios.

References

[1] G. A. Van Kleef, A. Cheshin, A. H. Fischer, and I. K. Schneider, "Editorial: The social nature of emotions," *Front Psychol*, vol. 7, no. JUN, Jun. 2016, doi: 10.3389/fpsyg.2016.00896.

[2] G. A. Van Kleef and S. Côté, "The Social Effects of Emotions," *Annual Review ofPsychology*, p. 629, Jul. 2021, doi: 10.1146/annurev-psych-020821.

[3] R. Hasson Marques, V. Violant-Holz, and E. Damião da Silva, "Emotions and decision-making in boardrooms—a systematic review from behavioral strategy perspective," *Front Psychol*, vol. 15, 2024, doi: 10.3389/fpsyg.2024.1473175.

[4] A. Shukla, "Utilizing AI and Machine Learning for Human Emotional Analysis through Speech-to-Text Engine Data Conversion," *Journal of Artificial Intelligence & Cloud Computing*, pp. 1–4, Mar. 2022, doi: 10.47363/JAICC/2022(1)145.

[5] Y. Zhou, X. Liang, Y. Gu, Y. Yin, and L. Yao, "Multi-Classifier Interactive Learning for Ambiguous Speech Emotion Recognition," Dec. 2020, doi: 10.1109/TASLP.2022.3145287.

[6] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3ER: Multiplicative Multimodal Emotion Recognition Using Facial, Textual, and Speech Cues," Nov. 2019, [Online]. Available: http://arxiv.org/abs/1911.05659

[7] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[8] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech Emotion Recognition with Dual-Sequence LSTM Architecture," Oct. 2019, doi: 10.1109/ICASSP40776.2020.9054629.

[9] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang, and B. Schuller, "Attention-enhanced connectionist temporal classification for discrete speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2019, pp. 206–210. doi: 10.21437/Interspeech.2019-1649.

[10] E. Morais, R. Hoory, W. Zhu, I. Gat, M. Damasceno, and H. Aronowitz, "Speech Emotion Recognition Using Self-Supervised Features," 2022.

[11] G. Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.06022

[12] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Commun*, vol. 120, pp. 11–19, Jun. 2020, doi: 10.1016/j.specom.2020.03.005.

[13] S. B. H. Avro, T. Taher, and N. Mamun, "EmoTech: A Multi-modal Speech Emotion Recognition Using Multi-source Low-level Information with Hybrid Recurrent Network," Jan. 2025, [Online]. Available: http://arxiv.org/abs/2501.12674

[14] W. Wu, C. Zhang, and P. C. Woodland, "Emotion recognition by fusing time synchronous and time asynchronous representations," Oct. 2020, doi: 10.1109/ICASSP39728.2021.9414880.

[15] J. He, X. Shi, X. Li, and T. Toda, "MF-AED-AEC: Speech Emotion Recognition by Leveraging Multimodal Fusion, Asr Error Detection, and Asr Error Correction," Jan. 2024, [Online]. Available: http://arxiv.org/abs/2401.13260

[16] Y. Lee, S. Yoon, and K. Jung, "Multimodal Speech Emotion Recognition using Cross Attention with Aligned Audio and Text," Jul. 2022, doi: 10.21437/Interspeech.2020-2312.

[17] K. M. Ibrahim, A. Perzo, and S. Leglaive, "Towards Improving Speech Emotion Recognition Using Synthetic Data Augmentation from Emotion Conversion," *International Conference on Acoustics, Speech, and Signal Processing*, 2024, doi: 10.1109/icassp48485.2024.10445740ï.

[18] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," Apr. 2019, doi: 10.21437/Interspeech.2019-2680.

[19] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.11477

[20] Z. Huijuan, Y. Ning, and W. Ruchuan, "Coarse-to-Fine Speech Emotion Recognition Based on Multi-Task Learning," *J Signal Process Syst*, vol. 93, no. 2–3, pp. 299–308, Mar. 2021, doi: 10.1007/s11265-020-01538-x.

[21] M. S. Fahad, A. Deepak, G. Pradhan, and J. Yadav, "DNN-HMM-Based Speaker-Adaptive Emotion Recognition Using MFCC and Epoch-Based Features," *Circuits Syst Signal Process*, vol.40(1): 466–489, Jan. 2021, doi:10.1007/s00034-020-01486-8.

[22] M. Xu, F. Zhang, and S. U. Khan, "Improve Accuracy of Speech Emotion Recognition with Attention Head Fusion," in *2020 10th Annual Computing and Communication Workshop and Conference, CCWC 2020*, Institute of Electrical and Electronics Engineers Inc., Jan. 2020, pp. 1058–1064. doi: 10.1109/CCWC47524.2020.9031207.

[23] Y. Wang *et al.*, "Multimodal transformer augmented fusion for speech emotion recognition," *Front Neurorobot*, vol. 17, 2023, doi: 10.3389/fnbot.2023.1181598.

[24] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04635

[25] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," 2007.

[26] 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," 2018, doi: 10.5281/zenodo.1188976.

[27] K. Fan *et al.*, "Neural Zero-Inflated Quality Estimation Model For Automatic Speech Recognition System," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01289