International Journal on Advanced Science Engineering Information Technology

Vol.14 (2024) No. 3
ISSN: 2088-5334

Text and Sound-Based Feature Extraction and Speech Emotion Classification for Korean

Jaechoon Jo [a], Soo Kyun Kim [b], Yeo-chan Yoon [c,*]

[a] *Department of Computer Education, Jeju National University, 63243 Jeju, Republic of Korea*
[b] *Department of Computer Engineering, Jeju National University, 63243 Jeju, Republic of Korea*
[c] *Department of Artificial Intelligence, Jeju National University, 63243 Jeju, Republic of Korea*
*Corresponding author: *ycyoon@jejunu.ac.kr*

***Abstract*— Embracing the complexities of human emotions conveyed through speech, this study ventures into Speech Emotion Recognition (SER) within the human-computer interaction domain, leveraging cutting-edge artificial intelligence technologies. Focusing on the auditory attributes of speech, such as tone, pitch, and rhythm, the research introduces an innovative approach that amalgamates deep learning techniques with the A Learnable Frontend for Audio Classification (LEAF) algorithm and wav2vec 2.0 pre-trained on a large corpus, specifically targeting Korean voice samples. This methodology underlines the capacity of these technologies to process and decipher complex vocal expressions, aiming to elevate emotion classification precision notably. The exploration extends the horizons of SER by accentuating auditory emotion cues and aspires to enrich machine interactions to be more intuitive and empathetic across various applications like healthcare and customer service. The outcomes underscore the efficacy of transformer-based models, particularly wav2vec 2.0 and LEAF, in capturing the subtle emotional states expressed in speech, thereby affirming the importance of auditory cues over conventional visual and textual indicators. The study's implications for further research herald a promising trajectory for evolving AI systems adept at nuanced emotion detection, thereby forging pathways toward more natural and human-centric interactions between individuals and machines. This advancement is crucial for developing empathetic AI that can seamlessly integrate into our daily lives, understanding and reacting to human emotions in a way that mirrors human understanding and compassion.**

***Keywords*— Speech emotion recognition; emotion detection; feature extraction; human-computer interaction.**

Manuscript received 8 Oct. 2023; revised 12 Dec. 2023; accepted 12 Mar. 2024. Date of publication 30 Jun. 2024.

I. INTRODUCTION

In human interaction, emotions serve as fundamental signals that shape our communication and understanding of one another. Among the various means through which emotions are expressed and perceived, speech is particularly potent, carrying the subtle nuances of emotional states through tonal variations, intensity, and rhythm. The field of human-computer interaction (HCI) has recognized the value of decoding these emotional signals from speech data, spurred by the rapid advancements in artificial intelligence (AI) technologies. As AI systems become increasingly integrated into daily life, accurately recognizing and responding to human emotions becomes essential for creating more intuitive and empathetic machine interactions. This has led to the emergence of speech emotion recognition (SER) technology as a critical area of research aimed at enhancing the quality of interactions between humans and machines by facilitating accurate emotion classification and enabling machines to interpret user intentions better.

The importance of SER technology stretches across numerous applications, from healthcare, where it can aid in diagnosing and treating emotional disorders, to customer service robots that can adapt their responses based on the user's emotional state. The potential of SER to revolutionize these interactions has sparked a growing interest among researchers, leading to a surge in studies exploring deep learning techniques for emotion recognition. With its ability to process and learn from vast amounts of data, deep learning has already shown remarkable success in fields such as image and speech recognition. Its application to SER promises to unlock new levels of accuracy and efficiency in emotion detection, moving beyond the limitations of traditional analytical methods.

Jo et al. [1] present an emotion recognition model that combines bidirectional long-short-term memory (Bi-LSTM) and convolutional neural networks (CNNs) using a Korean speech emotion database to improve accuracy in human-computer interaction.

Wagner et al. [2] present an in-depth analysis of transformer-based models, specifically wav2vec 2.0 and HuBERT, for speech emotion recognition (SER). It focuses on the performance of these models across the dimensions of arousal, dominance, and valence, highlighting their robustness and fairness in gender representation. The study demonstrates that transformer architectures, even without explicit linguistic information, can significantly improve valence prediction, implicitly offering insights into the models' ability to learn linguistic cues from speech data.

Hema et al. [3] introduce a Speech Emotion Recognition (SER) system that utilizes Convolutional Neural Networks (CNN) and Mel-frequency Cepstral Coefficients (MFCC) for feature extraction to identify underlying emotions in speech signals. This method outperforms existing systems by achieving higher accuracy and lower false favorable rates, emphasizing the effectiveness of spectral and prosodic feature utilization in emotion detection.

Min et al. [4] propose an innovative Hate Speech Detection (HSD) method named EHSor, which integrates emotion detection to improve HSD performance. The study employs Multi-Label Learning (MLL) and Multi-Task Learning (MTL) frameworks to enhance the detection process by exploring the correlation between hate speech and negative emotional states. The method leverages a shared BERT encoder and employs pseudo-multi-label data to train the system, effectively capturing the complex relationship between hate speech and emotional states and providing a novel approach to addressing HSD challenges.

Singh et al. [5] introduce a novel gender-dependent training approach for improving the accuracy of Speech Emotion Recognition (SER) systems. It focuses on leveraging Mel-frequency Cepstral Coefficients (MFCC) and their variants, utilizing a Convolutional Neural Network (CNN) architecture specifically tailored to recognize emotions in speech based on the speaker's gender. This method aims to enhance human-machine interaction by providing a more nuanced understanding of emotional expressions in speech.

Kumar et al. [6] present a speech-emotion recognition system using a multilayer neural network for smart assistance. This system can detect a range of human emotions from voice messages, including worry, surprise, sadness, happiness, hate, and love. It utilizes voice processing techniques to trigger actions like alerts through buzzers and LEDs based on the identified emotion, aiming to enhance human-machine interaction and support in various environments such as households and hospitals.

Jain et al. [7] detail the use of a Support Vector Machine (SVM) for Speech Emotion Recognition (SER), aiming to classify speech into four emotions: sadness, anger, fear, and happiness. It highlights the importance of feature extraction, utilizing Mel-frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding Coefficients (LPCC) alongside pitch, energy, and speech rate analysis. The research compares two classification strategies: One against All (OAA) and Gender-Dependent Classification, emphasizing the role of SVM in improving emotion recognition accuracy through sophisticated feature analysis and classification techniques.

Kadiri et al. [8] introduce an automatic emotion detection system from speech, emphasizing the significant role of excitation features extracted around glottal closure instants (GCIs). It proposes using the Kullback-Leibler distance to measure deviations between emotional and neutral speech, demonstrating the importance of excitation features in distinguishing between these speech types. This approach is novel in its minimal reliance on training data and language-independent applicability, offering a new direction for emotion detection research.

Aouani et al. [9] introduce a novel gender-dependent training approach for improving the accuracy of Speech Emotion Recognition (SER) systems. It focuses on leveraging Mel-frequency Cepstral Coefficients (MFCC) and their variants, utilizing a Convolutional Neural Network (CNN) architecture specifically tailored to recognize emotions in speech based on the speaker's gender. This method aims to enhance human-machine interaction by providing a more nuanced understanding of emotional expressions in speech.

Issa et al. [10] introduce a novel gender-dependent training approach for improving the accuracy of Speech Emotion Recognition (SER) systems. It focuses on leveraging Mel-frequency Cepstral Coefficients (MFCC) and their variants, utilizing a Convolutional Neural Network (CNN) architecture specifically tailored to recognize emotions in speech based on the speaker's gender. This method aims to enhance human-machine interaction by providing a more nuanced understanding of emotional expressions in speech.

Akçay et al. [11] overview the methodologies, technologies, and challenges in Speech Emotion Recognition (SER). It discusses various aspects of SER, including emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. The survey highlights the evolution of SER over two decades, emphasizing recent advancements and the integration of deep learning techniques. It provides a detailed examination of current literature, identifying gaps and suggesting future research directions.

Emotion recognition research has historically concentrated on visual and textual cues, such as facial expressions and linguistic patterns [12]–[21]. However, the auditory dimension of speech offers a rich vein of emotional information that remains relatively untapped. Vocal attributes like tone, pitch, and speaking rate can convey a broad spectrum of emotions, from joy and surprise to anger and sadness, often with greater subtlety and complexity than visual cues alone. Recognizing the potential of these vocal characteristics, our research seeks to advance the field of SER by focusing on the auditory aspects of emotion recognition.

Our study introduces an innovative approach that combines the latest advancements in deep learning with the unique capabilities of the LEAF algorithm [22]. This learnable audio frontend combines Gabor filters, learnable pooling, sPCEN for audio signal processing, and wav2vec [23]. Pretraining on large datasets has recently brought significant success to deep learning approaches [24]–[31]. Wav2vec 2.0 is a self-supervised learning framework for speech recognition. It pretrains on unlabeled audio data to learn speech representations without manual transcription. By explicitly

targeting Korean voice samples, we explore new methodologies for classifying emotions through speech, leveraging the strengths of these algorithms to process and analyze the intricate patterns of vocal expressions. This approach broadens the scope of SER and offers a more nuanced understanding of how emotions are communicated and perceived through speech. Our research contributes to the ongoing development of SER technologies, aiming to bridge the gap between human emotions and machine interpretation. By enhancing the ability of AI systems to recognize and respond to emotional cues in speech, we move closer to creating more natural, intuitive, and human-centric interactions between people and machines.

II. Materials and Method

A. Preliminaries

(LEAF) LEAF (Learnable Frontend for Audio Classification) is an entirely learnable architecture designed to replace mel-filterbanks in audio classification tasks. By training on diverse audio signals, including speech, music, and environmental sounds, the authors demonstrate that LEAF outperforms traditional mel-filterbanks and previous learnable alternatives across a wide range of audio classification tasks. The key innovation is the end-to-end learnability of all operations in audio feature extraction, including filtering, pooling, compression, and normalization, which enables superior performance with significantly fewer parameters compared to the state-of-the-art. This paper employs the LEAF model to extract voice signal features, including pitch, tone, and volume.

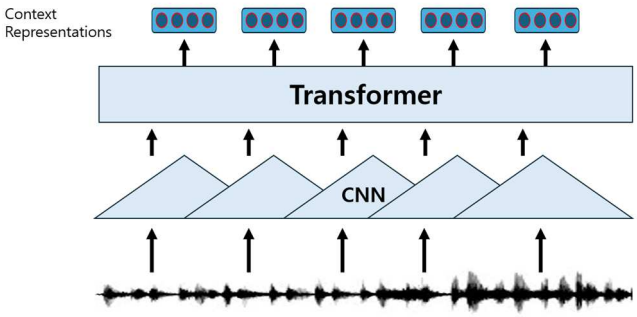

Fig. 1 Architecture of Wav2Vec 2.0

Wav2Vec 2.0 is a framework for self-supervised learning of speech representations that significantly advances the field of automatic speech recognition (ASR). It introduces a novel approach that masks latent representations of raw waveform and solves a contrastive task over quantized speech representations, demonstrating substantial improvements in word error rates (WER) on the Librispeech dataset, especially in low-resource scenarios. The framework outperforms the best semi-supervised methods and previous state-of-the-art results by leveraging large amounts of unlabeled data for pre-training, followed by fine-tuning a small amount of labeled data. This method shows promise for developing ASR systems with minimal labeled data, potentially expanding the accessibility of speech recognition technology across various languages and dialects. Fig 1 shows the architecture of Wav2Vec 2.0. The bottom of the diagram shows the input raw audio waveform. The next layer up, labeled 'CNN', represents a series of 1-dimensional convolutional neural networks that process the raw audio signal to extract features over time. These features are then passed up to the 'Transformer', the core of the wav2vec 2.0 model. The Transformer is a type of neural network that processes sequential data and is known for its effectiveness in natural language processing tasks. At the top, 'Context Representations' are the output from the Transformer. These representations are contextually rich, considering the immediate acoustic features and the broader context within the audio sequence. These context representations can then be used for various downstream tasks, such as speech recognition, by adding an appropriate head to the pre-trained model. Yi [32] explores the application of wav2vec 2.0, pre-trained on English speech, to low-resource Automatic Speech Recognition (ASR) tasks across six languages. It significantly improves over previous models, highlighting its adaptability to linguistic contexts and its efficiency with coarse-grained modeling units. The study underscores wav2vec 2.0's potential in handling real-world, low-resource speech recognition challenges, showing over 20% relative improvements in six languages and particularly notable gains in English. In this paper, we utilize wav2vec to extract feature vectors from Korean voices for classifying emotions based on voice. Recently, various works have attempted to employ wav2vec to extract features for emotion recognition tasks. Mohamed et al. [33] introduce an advanced deep learning model to identify emotions in Arabic speech, utilizing state-of-the-art audio representation technologies, wav2vec2.0 and HuBERT, on the Arabic BAVED audio dataset. The study showcases wav2vec2.0's ability to achieve an impressive accuracy of 89% through various experiments. The research highlights the effectiveness of these models in recognizing emotional states from speech, demonstrating significant advancements in Arabic speech emotion recognition. This work sets a new benchmark and opens pathways for future research to enhance feature sets and expand dataset size for broader recognition tasks. Sharma et al. [34] present a Multi-Lingual (MLi) and Multi-Task Learning (MTL) system for Speech Emotion Recognition (SER), leveraging the pre-trained wav2vec 2.0 model. It fine-tunes this model on twenty-five open-source datasets across 13 languages and 7 emotion categories. Key findings include the model's superior performance over Pre-trained Audio Neural Network (PANN) models, achieving up to 8.6% improvement in specific tasks. The MTL approach notably enhances the system's accuracy across different languages, showcasing the model's robustness and adaptability in multi-lingual emotion recognition tasks. In this paper, we employ Wav2Vec as a language-related feature extractor. Using a pre-trained feature extractor, textual and voice information can be encoded.

B. Text and Sound-Based Feature Extraction and Emotion Classification

Fig 2 illustrates the architecture proposed for the classification of emotions using vocal inputs. Initially, the system processes the raw audio signal, transforming it into features using a pre-trained Wav2Vec model. Then, after this initial transformation, the features are sent to the LEAF (Learnable Audio Frontend) model to be transformed into a short-term power spectrum representation. This step is crucial

as it captures the temporal dynamics and frequency components essential for interpreting human emotions from speech. The synergy between the learnable frontend and the Wav2Vec model forms the cornerstone of our methodology, allowing for the integration of signal characteristics with linguistic features. By doing so, the system gains the capacity to discern emotional subtleties in the human voice, which could be significantly influenced by language and culture.

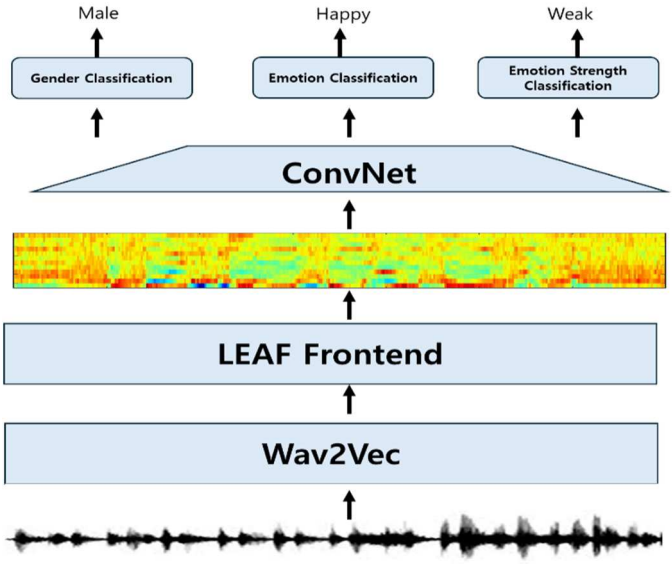

Fig. 2 Architecture of the Proposed System

C. *Multi-task Learning for Audio Classification.*

Multi-task learning has made significant strides in enhancing the robustness of predictive models [35]. By training a model on multiple tasks simultaneously, the shared representations can lead to more generalized features and prevent overfitting on a single task. Our approach leverages this paradigm by implementing a multi-task classification framework that incorporates various attributes of the audio signal. The dataset used for training encapsulates several variables, including but not limited to gender identification, emotion intensity, and the emotional state itself. This dataset enables the model to perform a comprehensive multi-class classification, which is beneficial in distinguishing subtle differences between emotional states. The uppermost three layers in Fig 2 represent the linear layers designated for multi-task learning. Each layer is responsible for a different classification task – gender, emotion, and emotion intensity. The interaction among these tasks is designed to reinforce the model's discriminative power. The gender classification layer focuses on identifying the speaker's gender, which can influence emotion perception in speech. The emotion classification layer categorizes the speaker's emotional state, which is the system's primary focus. Finally, the emotion strength classification layer gauges the intensity of the expressed emotion, which can vary from subtle to intense. The multi-task learning framework allows the shared representation to be fine-tuned by the gradients derived from each task, ensuring that the layers capture universally applicable features across all tasks. Moreover, this framework benefits from the interrelatedness of the functions. For example, recognizing gender may provide helpful context for interpreting emotion, as cultural norms can lead to different emotional expressions among genders. To train this model, we employ a joint loss function that combines the losses from each task, allowing us to optimize the network cohesively. This streamlines the training process and aligns the model's objectives towards a common goal: to understand and classify the emotional content in speech accurately. The implications of such a system are far-reaching, especially for languages like Korean, where emotional expressions can be highly context-dependent and influenced by social hierarchies. By tailoring the model to recognize the nuances of Korean speech, we aim to achieve high accuracy in emotion recognition, which can be applied to various fields such as interactive voice response systems, mental health assessment, and human-computer interaction. In conclusion, our proposed system integrates the learnable LEAF frontend with a sophisticated wav2vec model, followed by a robust multi-task learning framework to classify gender, emotion, and emotion intensity from speech. This comprehensive approach enhances the model's generalizability across tasks and caters to the intricate variances found in Korean speech, making it a significant step forward in speech emotion classification.

III. Results and Discussion

A. *Dataset*

In this study, we developed a comprehensive Korean speech emotion dataset by engaging 120 voice actors with notable experience in various domains, including animations, dramas, and web-based audio content. These actors were assigned to vocalize texts, with each text's emotional tone aligning with the predefined categories of an emotion-labeled dataset. This dataset comprises sentences randomly sourced from the internet, categorized based on their emotional content. Through this methodology, each participating voice actor contributed to the dataset by recording 500 to 600 sentences, culminating in a robust compilation of 68,000 data points. A distinctive feature of this dataset is the inclusion of onomatopoeia and exclamations, accounting for 5% of each actor's contributions, thereby enhancing the dataset's emotional diversity. Specific sounds such as laughter, crying, and shouting were incorporated into categories corresponding to happiness, sadness, and anger. The dataset is structured into four primary emotional classes—joy, sadness, neutral, and anger—each containing 17,000 utterances. For analytical and developmental purposes, these utterances were subdivided into training (13,000 utterances), validation (2,000 utterances), and testing (2,000 utterances) sets, facilitating a structured approach to model training and performance evaluation. Additionally, the dataset was meticulously curated to consider the gender and age distribution of the voice actors. The gender distribution was balanced at a 50% male to 50% female ratio, with the age distribution comprising 40% in their 20s, 40% in their 30s, and 20% aged 40 and above. Furthermore, each utterance was categorized into four intensity levels, providing a nuanced understanding of emotional expression.

B. *Experimental Result.*

We experimented to evaluate the proposed method. Four models were compared:

- Wav2Vec with Multitask Learning: This model was trained with the architecture proposed in Fig. 2. It includes Wav2Vec feature extraction and a LEAF frontend model to transform the input voice into features for a convolutional neural network (ConvNet). The model was trained with a multitask learning strategy.
- Wav2Vec without Multitask Learning: To identify the effectiveness of the multitask learning strategy, a Wav2Vec model was trained without it.
- LEAF only with Multitask Learning: To identify the effectiveness of the proposed model, which constructs a pipeline of pretrained Wav2Vec and a LEAF frontend, the proposed model was compared with a LEAF model trained with multitask learning.
- LEAF only without Multitask Learning: A LEAF model was also trained without multitask learning.

Fig 3 shows the loss curves of the baseline models and the proposed model. The proposed Wav2Vec with multitask learning shows the best loss curve. The proposed Wav2Vec and LEAF pipeline outperform the LEAF-only models. The multitask learning strategy shows better performance than single-task learning in both the Wav2Vec and LEAF pipeline and the LEAF-only model.

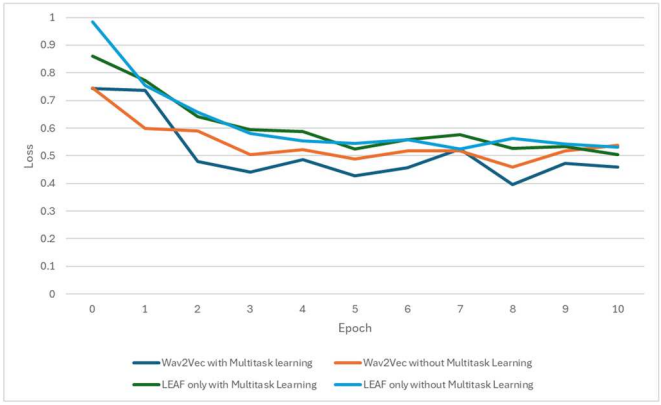

Fig. 3 Loss changes between models.

The performance of the models was evaluated using standard metrics for speech emotion classification tasks. Here, we present the results using accuracy as the primary metric. Accuracy represents the percentage of speech samples where the predicted emotion category matches the ground truth label.

TABLE I
MULTITASK CLASSIFICATION PERFORMANCE.

Model	Accuracy (%)
LEAF only without Multitask Learning	80.76
LEAF only with Multitask Learning	82.33
Wav2Vec without Multitask Learning	83.72
Wav2Vec with Multitask learning	86.40

Table 1 presents the accuracy for each model on the test set. The proposed Wav2Vec with multitask learning achieves the highest accuracy, outperforming all baseline models. This confirms the effectiveness of the proposed method in Korean speech emotion classification. The Wav2Vec without Multitask Learning model performs less than the proposed model with multitask learning. This difference highlights the benefit of incorporating the multitask learning strategy, potentially improving the model's ability to learn generalizable features applicable to emotion classification. The LEAF-only models consistently show lower performance with and without multitask learning than the Wav2Vec-based models. This observation suggests that Wav2Vec features capture emotional cues crucial for accurate classification, and the LEAF architecture alone might not be sufficient for optimal performance on this specific task. The LEAF only with the Multitask Learning model performs slightly better than the LEAF only without the Multitask Learning model. While the improvement is modest, it indicates that multitasking learning can potentially benefit even models that do not leverage Wav2Vec features.

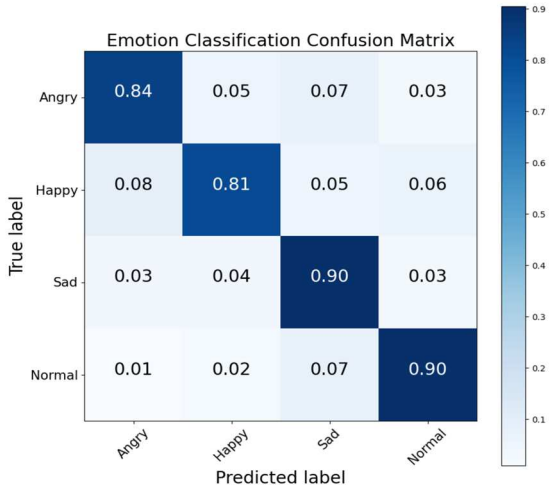

Fig. 4 Emotion Classification Confusion Matrix of proposed model

The confusion matrix, as illustrated in Fig 2, provides a visual representation of the performance of our speech emotion classification system for Korean language datasets. The matrix reveals that the 'Sad' emotion classification outperforms other emotional states, with a substantial true positive rate of 90%. This high accuracy can be attributed to distinct acoustic features effectively captured by the feature extraction in the case of 'Sad' vocal expressions. In contrast, the 'Happy' emotion is identified with the lowest accuracy. This is possibly due to the nuanced variations in the 'Happy' vocal expressions that may be similar to those found in 'Angry' and 'Normal' states, leading to a more distributed pattern of misclassification among these categories. The system's tendency to confuse 'Happy' with 'Angry' and 'Normal' suggests a significant overlap in the feature space for these emotions. The performance metrics for 'Angry' and 'Normal' emotions are satisfactory, achieving a classification accuracy of 84% and 90%, respectively. Notably, 'Normal' emotion, which we may consider as a baseline or control, is classified with high precision, indicating the system's effectiveness in distinguishing neutral emotional states from more expressive ones.

Future work will focus on improving the 'Happy' emotion classification by exploring additional feature extraction techniques and considering a more extensive and more varied dataset to train the model. The overarching goal is to enhance the emotional granularity of the classification system, ensuring that each emotional state is represented with high fidelity in the automated recognition process. Moreover,

integrating contextual information and linguistic cues into the feature extraction process may offer a pathway to improved classification accuracy across all emotions.

C. Training Details

For hyperparameter tuning, we employed a grid search approach to identify the optimal configuration for each model. We set the batch size to 128 and used a learning rate of 5e-3 with 10 epochs.

IV. Conclusion

This research has demonstrated substantial improvement in the precision of emotion classification, particularly for the Korean language. The combination of text and sound-based feature extraction methodologies has proven effective in capturing the nuanced expressions of emotions in speech, which is crucial for enhancing human-computer interactions in various applications, including healthcare and customer service. The study's dataset, consisting of voice recordings from 120 actors and 68,000 data points, underscores the comprehensive approach to understanding emotional expressions in Korean speech. Experimental results showcasing the superiority of the wav2vec model with multitask learning in accuracy further validate the proposed method's effectiveness. Future work should focus on refining these models and exploring additional linguistic and cultural nuances to further extend SER's applicability and accuracy. This research contributes to the SER field and opens new pathways for creating more empathetic and intuitive AI systems capable of understanding and interacting with human emotions more naturally.

Acknowledgment

This work was supported by a research grant from Jeju National University in 2024

References

[1] A.-H. Jo and K.-C. Kwak, “Speech Emotion Recognition Based on Two-Stream Deep Learning Model Using Korean Audio Information,” Applied Sciences, vol. 13, no. 4, p. 2167, Feb. 2023, doi:10.3390/app13042167.

[2] J. Wagner et al., “Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 9, pp. 10745–10759, Sep. 2023, doi: 10.1109/tpami.2023.3263585.

[3] C. Hema and F. P. Garcia Marquez, “Emotional speech Recognition using CNN and Deep learning techniques,” Applied Acoustics, vol. 211, p. 109492, Aug. 2023, doi: 10.1016/j.apacoust.2023.109492.

[4] C. Min et al., “Finding hate speech with auxiliary emotion detection from self-training multi-label learning perspective,” Information Fusion, vol. 96, pp. 214–223, Aug. 2023, doi:10.1016/j.inffus.2023.03.015.

[5] V. Singh and S. Prasad, “Speech emotion recognition system using gender dependent convolution neural network,” Procedia Computer Science, vol. 218, pp. 2533–2540, 2023, doi:10.1016/j.procs.2023.01.227.

[6] Kumar et al., “Multilayer Neural Network Based Speech Emotion Recognition for Smart Assistance,” Computers, Materials & Continua, vol. 74, no. 1, pp. 1523–1540, 2023, doi:10.32604/cmc.2023.028631.

[7] Jain, Manas, et al. "Speech emotion recognition using support vector machine." arXiv preprint arXiv:2002.07590, 2020.

[8] S. R. Kadiri and P. Alku, “Excitation Features of Speech for Speaker-Specific Emotion Detection,” IEEE Access, vol. 8, pp. 60382–60391, 2020, doi: 10.1109/access.2020.2982954.

[9] H. Aouani and Y. B. Ayed, “Speech Emotion Recognition with deep learning,” Procedia Computer Science, vol. 176, pp. 251–260, 2020, doi: 10.1016/j.procs.2020.08.027.

[10] D. Issa, M. Fatih Demirci, and A. Yazici, “Speech emotion recognition with deep convolutional neural networks,” Biomedical Signal Processing and Control, vol. 59, p. 101894, May 2020, doi:10.1016/j.bspc.2020.101894.

[11] M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” Speech Communication, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.

[12] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. M. Sadeeq, and S. Zeebaree, “Multimodal Emotion Recognition using Deep Learning,” Journal of Applied Science and Technology Trends, vol. 2, no. 01, pp. 73–79, May 2021, doi: 10.38094/jastt20291.

[13] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, “Comparing Recognition Performance and Robustness of Multimodal Deep Learning Models for Multimodal Emotion Recognition,” IEEE Transactions on Cognitive and Developmental Systems, vol. 14, no. 2, pp. 715–729, Jun. 2022, doi: 10.1109/tcds.2021.3071170.

[14] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, “Cross-Subject Multimodal Emotion Recognition Based on Hybrid Fusion,” IEEE Access, vol. 8, pp. 168865–168878, 2020, doi:10.1109/access.2020.3023871.

[15] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 02, pp. 1359–1367, Apr. 2020, doi:10.1609/aaai.v34i02.5492.

[16] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “EmotiCon: Context-Aware Multimodal Emotion Recognition Using Frege’s Principle,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2020, doi:10.1109/cvpr42600.2020.01424.

[17] C. Luna-Jiménez, D. Griol, Z. Callejas, R. Kleinlein, J. M. Montero, and F. Fernández-Martínez, “Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning,” Sensors, vol. 21, no. 22, p. 7665, Nov. 2021, doi: 10.3390/s21227665.

[18] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, and D. Zhang, “Multimodal Emotion Recognition With Temporal and Semantic Consistency,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3592–3603, 2021, doi: 10.1109/taslp.2021.3129331.

[19] S. Lee, D. K. Han, and H. Ko, “Multimodal Emotion Recognition Fusion Analysis Adapting BERT With Heterogeneous Feature Unification,” IEEE Access, vol. 9, pp. 94557–94572, 2021, doi:10.1109/access.2021.3092735.

[20] S. S. R, J. S. B, and R. R, “Comprehensive Speech Emotion Recognition System Employing Multi-Layer Perceptron (MLP) Classifier and libRosa Feature Extraction,” 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Nov. 2023, doi: 10.1109/icscna58489.2023.10370394.

[21] Y. C. Yoon, “Can We Exploit All Datasets? Multimodal Emotion Recognition Using Cross-Modal Translation,” IEEE Access, vol. 10, pp. 64516–64524, 2022, doi: 10.1109/access.2022.3183587.

[22] Zeghidour, Neil, et al. "LEAF: A learnable frontend for audio classification." arXiv preprint arXiv:2101.08596, 2021.

[23] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 1044, 12449–12460.

[24] Zoph, Barret, et al. "Rethinking pre-training and self-training." In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 323, 3833–3845.

[25] F.-L. Chen et al., “VLP: A Survey on Vision-language Pre-training,” Machine Intelligence Research, vol. 20, no. 1, pp. 38–56, Jan. 2023, doi: 10.1007/s11633-022-1369-5.

[26] Bao, Hangbo, et al. "Beit: Bert pre-training of image transformers." arXiv preprint arXiv:2106.08254, 2021.

[27] El-Nouby, Alaaeldin, et al. "Are large-scale datasets necessary for self-supervised pre-training?." arXiv preprint arXiv:2112.10740, 2021.

[28] Jiang, Ziyu, et al. "Robust pre-training by adversarial contrastive learning." In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 1359, 16199–16210.

[29] L. H. Li et al., “Grounded Language-Image Pre-training,” 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022, doi: 10.1109/cvpr52688.2022.01069.

[30] L. Ruan and Q. Jin, “Survey: Transformer based video-language pre-training,” AI Open, vol. 3, pp. 1–13, 2022, doi:10.1016/j.aiopen.2022.01.001.

[31] W. Ahmad, S. Chakraborty, B. Ray, and K.-W. Chang, “Unified Pre-training for Program Understanding and Generation,” Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, doi: 10.18653/v1/2021.naacl-main.211.

[32] Yi, Cheng, et al. "Applying wav2vec2. 0 to speech recognition in various low-resource languages." arXiv preprint arXiv:2012.12121, 2020.

[33] Mohamed, Omar, and Salah A. Aly. “ASER: Arabic Speech Emotion Recognition Employing Wav2vec2.0 and HuBERT Based on BAVED Dataset,” Transactions on Machine Learning and Artificial Intelligence, vol. 9, no. 6, pp. 1–8, Nov. 2021, doi: 10.14738/tmlai.96.11039.

[34] M. Sharma, “Multi-Lingual Multi-Task Speech Emotion Recognition Using wav2vec 2.0,” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2022, doi: 10.1109/icassp43922.2022.9747417.

[35] Y. C. Yoon, S. Y. Park, S. M. Park, and H. Lim, “Image classification and captioning model considering a CAM-based disagreement loss,” ETRI Journal, vol. 42, no. 1, pp. 67–77, Jul. 2019, doi:10.4218/etrij.2018-0621.