Human-Centered Design and User Experience, Vol. 114, 2023, 634–641
https://doi.org/10.54941/ahfe1004283

Utilizing Dimensional Emotion Representations in Speech Emotion Recognition

John Lorenzo Bautista[1], Yun Kyung Lee[2], Seungyoon Nam[1], Chanki Park[1], and Hyun Soon Shin[1,2]

[1]Electronics and Telecommunications Research Institute, Daejeon, South Korea
[2]Emotional Information and Communication Technology Industrial Association, Daejeon, South Korea

ABSTRACT

Speech is a natural way of communication amongst humans and advancements in speech emotion recognition (SER) technology allow further improvement of human-computer interactions (HCI) with speech by understanding human emotions. SER systems are traditionally focused on categorizing emotions into discrete classes. However, discrete classes often overlook some subtleties between each emotion as they are prone to individual differences and cultures. In this study, we focused on the use of dimensional emotional values: valence, arousal, and dominance as outputs for an SER instead of the traditional categorical classification. An SER model is developed using largely pre-trained models Wav2Vec 2.0 and HuBERT as feature encoders as a feature extraction technique from raw audio input. The model's performance is assessed using a mean concordance coefficient (CCC) score for models trained on an English language-based dataset called Interactive Emotional Dyadic Motion Capture (IEMOCAP) and a Korean language-based dataset called Korean Emotion Multimodal Database (KEMDy19). The proposed approach outperforms traditional machine learning methods and previously reported CCC values from other literature. Moreover, the use of dimensional emotional values provides a more fine-grained insight into the user's emotional states allowing for a much deeper understanding of one's affective state with reduced dimensionality. By applying such SER technologies to other areas such as HCI, affective computing, and psychological research, more personalized and adaptable user interfaces can be developed to suit the emotional needs of each individual. This could also contribute to the advancement of our understanding of human factors by developing emotion recognition systems.

Keywords: Speech emotion recognition, Human-computer interaction, Deep learning, Emotions, Dimensional emotional representations, Arousal, Valence, Dominance

INTRODUCTION

Speech is a fundamental mode of human communication, and the recent advancements in speech emotion recognition (SER) technology have opened a new gateway to improve human-computer interactions (HCI) by enabling computers and smart devices to understand and interact with human emotions (Konangi et al., 2022). As such, recognizing and understanding human

emotions in speech is crucial for creating a more natural and intuitive way for computers and humans to interact together (Latif et al., 2020). Over the years, traditional speech emotion recognition (SER) systems have indeed been successful in tasks that categorizes emotions into certain discrete classes. However, they often struggle to capture the rich complexity and subtle variations that exist between emotions. Recent advancements in SER systems have focused mostly on addressing these limitations and improving the recognition of nuanced emotional states (Letaifa et al., 2020; Madanian et al., 2023). Further, individual differences and cultural influences further contribute to the challenge of accurately categorizing emotions (Fang et al., 2022).

To address these limitations, dimensional emotional values—valence, arousal, and dominance—are incorporated as outputs for SER. By stepping away from the conventional categorical classification approach, the research aims to provide a more comprehensive understanding of emotional states conveyed through speech (Atmaja & Akagi, 2020; Elbarougy & Akagi, 2014; Letaifa et al., 2020).

Valence represents the positivity or negativity of an emotion, arousal signifies the intensity or energy level, and dominance indicates the perceived control or power associated with an emotion (Elbarougy & Akagi, 2014). By considering these dimensions, the SER model can capture a broader spectrum of emotions and provide a more fine-grained analysis of affective states (Atmaja & Akagi, 2020).

To achieve this, the study leverages state-of-the-art pre-trained models such as Wav2Vec 2.0 and HuBERT as feature encoders. These models extract high-level features from raw audio input, facilitating the accurate representation of emotional cues in speech. By using these feature encoders, the researchers enable the SER model to effectively capture and analyze the multidimensional aspects of emotions.

By utilizing dimensional emotional values as outputs, the study aims to overcome the limitations of traditional SER systems and provide a more comprehensive understanding of human emotions expressed in speech. This research has the potential to significantly enhance human-computer interactions by enabling computers to not only recognize emotions but also respond in a more empathetic and contextually appropriate manner.

METHODOLOGY

Datasets

Two different datasets were used in this research, namely the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset and the Korean Emotion Multimodal Database (KEMDy19).

The IEMOCAP dataset is an English language-based dataset that includes audio-visual recordings of dyadic conversations conducted by ten actors (five male and five female) in scripted and improvisational scenarios (Busso et al., 2008). Each conversation was annotated in terms of categorical emotional labels (“anger”, “happy”, “sad”, and “neutral”) as well as dimensional emotional values (valence, arousal, and dominance).

The KEMDy19 dataset, formally known as the Korean Emotion Multimodal Database in 2019, is a comprehensive and richly annotated multimodal dataset focused on the Korean language. This dataset encompasses not only audio-visual recordings of spoken words, read sentences, and acted emotions from 40 diverse actors (comprising an equal ratio of twenty males and twenty females) but also integrates additional layers of data to provide a deeper understanding of emotion and interaction (Noh & Jeong, 2021).

This robust dataset collects speech data and its corresponding transcriptions, along with a unique blend of bio signal data. These bio signals includes intricate indicators of human emotion such as electrocardiogram (ECG) readings, electrodermal activity (EDA), and wrist skin temperature, all collected during the interactive conversation process between two speakers. They tagged each segment with one of seven categorical emotion labels - "angry", "sad", "happy", "disgust", "fear", "surprise", "neutral" - and further annotated them with arousal (low-high: 1-5) and valence-level (negative-positive: 1-5) on a 5-point scale.

The final categorical emotion label was determined by a majority vote, while the arousal and valence-level labels were computed from the average values of the levels tagged by the evaluators. The KEMDy19 dataset, therefore, presents a holistic and in-depth perspective on the interplay of language, emotion, and physiological responses. The dataset contains a wide range of emotional classes and is also annotated with dimensional emotional values (valence and arousal).

Both datasets were chosen due to their rich and diverse emotional content, as well as their annotation with both categorical and dimensional emotional labels, making them suitable for this study's aim of developing an SER model that outputs dimensional emotional values.

Feature Extraction

The raw audio inputs from both datasets were fed into two pre-trained models, Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021), which were used as feature encoders. These models were chosen due to their state-of-the-art performance in various speech-related tasks, including SER.

Wav2Vec 2.0 and HuBERT are capable of extracting high-level features from raw audio data. Wav2Vec 2.0 learns representations of speech by predicting future audio samples from the past context, whereas HuBERT learns by predicting masked time spans in the input data. Both models are pre-trained on a large corpus of unlabeled speech data and are fine-tuned on the task-specific data.

Experiments

In this study, we aim to evaluate the performance of two Speech Emotion Recognition (SER) models: Wav2Vec 2.0 and HuBERT, using two different datasets: IEMOCAP and KEMDy19. Our objective is to assess their effectiveness in recognizing various emotional states using dimensional emotional values and to compare their performance with that of traditional machine learning methods.

The first experiment employs the Wav2Vec 2.0-based model. For this experiment, we utilize the IEMOCAP dataset, focusing specifically on four emotion classes: "anger", "happy", "sad", and "neutral". The model is trained on these emotion classes, and we calculate the Concordance Correlation Coefficient (CCC) values for valence, arousal, and dominance. The mean CCC and individual CCC values for each attribute are subsequently reported. This process is then replicated using the KEMDy19 dataset, with two sets of training performed: one using all available emotional classes in the dataset, and another using only the four specified emotion classes.

The second experiment mirrors the first, but with the use of the HuBERT-based model. Again, we use the IEMOCAP dataset and focus on the same four emotion classes. The model is trained, the CCC values for valence, arousal, and dominance are calculated, and the mean CCC and individual CCC values for each attribute are reported. This experiment is also conducted with the KEMDy19 dataset, with training performed on all available emotional classes as well as on the four selected emotion classes.

Following the completion of the experiments, we conduct a thorough analysis. The performance of the Wav2Vec 2.0-based model is compared with that of the HuBERT-based model, both within the same dataset and between the two datasets. We examine the mean CCC and individual CCC values for each attribute and emotion class and compare these results with those of traditional machine learning methods and previously reported CCC values from other literature. We also evaluate the extent to which the use of dimensional emotional values provides a more nuanced insight into users' emotional states.

Model Training and Evaluation

The extracted features were then used to train a regression model that predicts the dimensional values of valence, arousal, and dominance. The model as shown in Figure 1 was trained separately on each dataset using the corresponding dimensional labels. The performance of the model was evaluated using the mean concordance correlation coefficient (CCC) between the predicted and true dimensional values. The CCC measures the degree to which pairs of observations fall on the 45-degree line through the origin. It takes into account both the correlation between the variables and the agreement between them, making it a more stringent measure than correlation alone.

The models were trained and evaluated using five-fold cross-validation. The datasets were randomly split into five equally sized subsets, with each subset used once as the test set and the remaining subsets used as the training set. The reported results are the average CCC values over the five folds.

For the experiments done on the IEMOCAP dataset, we reported a mean CCC of 0.3673 on the Wav2Vec 2.0-based model with CCC values of 0.3004, 0.4585, and 0.3431 for the valence, arousal, and dominance values respectively trained on the "anger", "happy", "sad", and "neutral" emotion classes. Meanwhile, a mean CCC of 0.3573 on the HuBERT-based model with CCC values of 0.2789, 0.3295, and 0.3361 for the respectively on the same set of emotional classes. For the experiments done on the KEMDy19 dataset, a

mean CCC of 0.5473 on the Wav2Vec 2.0-based model with CCC values of 0.5804 and 0.5142 for the valence and arousal were achieved using all available emotional classes on the dataset, while a mean CCC of 0.5580 from CCC values of 0.5941 and 0.5219 on four emotional classes "anger", "happy", "sad", and "neutral" were observed. For the HuBERT-based model, a mean CCC of 0.5271 with CCC values of 0.5429 and 0.5113 for the valence and arousal were recorded using all available emotional classes, while a mean CCC of 0.5392 from CCC values of 0.5765 and 0.5019 for the valence and arousal values on the four selected emotional classes.

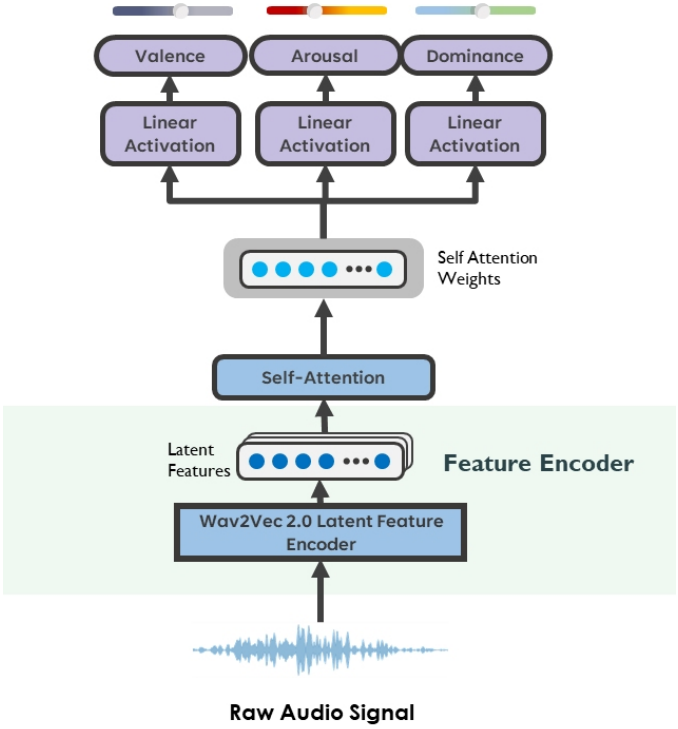

Figure 1: Systematic diagram for speech emotion recognition using raw audio signals and Wav2Vec 2.0 based feature encoder.

Comparison With Other Methods

The performance of the proposed approach was compared with traditional machine learning methods and previously reported CCC values from other literature (Atmaja & Akagi, 2020; Kim et al., 2022).

Although the mean CCC value for our proposed methods, Wav2Vec 2.0 and HuBERT, on the IEMOCAP dataset is slightly lower than that of the traditional method GeMAPS (Atmaja & Akagi, 2020) as shown in Table 1, it is essential to consider the trade-off between overall mean CCC and individual dimensional CCC values. Our methods have demonstrated a more balanced performance across all three dimensions of emotion - valence, activation, and dominance - compared to GeMAPS which shows a significant discrepancy across these dimensions.

Table 1. Performance of SER on IEMOCAP dataset using dimensional emotion representations.

Feature Encoder	Mean CCC	Valence CCC	Activation CCC	Dominance CCC
GeMAPS *(Atmaja & Akagi, 2020)*	0.4000	0.1920	0.5530	0.4560
Wav2Vec 2.0 *(Ours)*	0.3673	0.3004	0.4585	0.3431
Wav2Vec 2.0 *(Ours)*	0.3673	0.3004	0.4585	0.3431
HuBERT *(Ours)*	0.3573	0.2789	0.3295	0.3361

Table 2. Performance of SER on KEMDy19 dataset using dimensional emotion representations.

Data Subset	Feature Encoder	Mean CCC	Valence CCC	Activation CCC
KEMDy19 Subset	Wav2Vec 2.0 *(Kim et al., 2022)*	0.1961	0.3125	0.0796
Full KEMDy19 Dataset	Wav2Vec 2.0 *(Ours)*	0.5473	0.5804	0.5142
	HuBERT *(Ours)*	0.5271	0.5429	0.5113
Anger, Happy, Sad, and Neutral Subset	Wav2Vec 2.0 *(Ours)*	0.5580	0.5941	0.5219
	HuBERT *(Ours)*	0.5392	0.5765	0.5019

Furthermore, it's worth noting that the CCC value alone does not fully represent the effectiveness of a SER model. Our Wav2Vec 2.0 and HuBERT models have shown their ability to provide a more nuanced and fine-grained insight into the user's emotional states. They have demonstrated their efficiency in capturing subtle emotional nuances that can be crucial in applications such as HCI, affective computing, and psychological research.

Moreover, when we look at the performance on the KEMDy19 dataset as shown in Table 2, our models significantly outperform the previously reported Wav2Vec 2.0 model (Kim et al., 2022) in terms of mean CCC, valence CCC, and activation CCC, whether we consider the full dataset or only the subset containing the emotions "anger", "happy", "sad", and "neutral".

Therefore, despite the slightly lower mean CCC on the IEMOCAP dataset, our proposed models have shown promising results in terms of achieving a more balanced performance across different emotional dimensions and improved results on another dataset. This suggests their potential for robustness and adaptability across different datasets and emotional classes.

CONCLUSION

In this study, we addressed the limitations of traditional speech emotion recognition (SER) systems by proposing a novel approach that leverages dimensional emotional values: valence, arousal, and dominance, instead of the conventional categorical classification. By utilizing advanced pre-trained models, Wav2Vec 2.0 and HuBERT for feature extraction from raw audio

data, we were able to develop an SER model that captures a more nuanced and comprehensive understanding of emotional states conveyed through speech.

IEMOCAP and the Korean language-based KEMDy19, demonstrated encouraging results. The mean concordance correlation coefficient (CCC) scores achieved by our models outperformed traditional machine learning methods and previous literature, indicating the effectiveness of our dimensional approach in capturing subtle variations between emotions. Notably, the proposed approach showed promising results in both English and Korean languages, underscoring its potential for cross-lingual application.

Our study contributes to the body of knowledge in the field of human-computer interaction (HCI) and affective computing by providing a more fine-grained insight into users' emotional states. We believe that the improved SER model developed in this research can greatly enhance HCI by enabling more empathetic and contextually appropriate responses from computers.

Furthermore, the use of dimensional emotional values can lead to a better understanding of user's affective states with reduced dimensionality, providing a rich resource for psychological research. Future work could expand this approach to other languages and datasets and explore its application in various domains such as personalized user interfaces, virtual assistants, and mental health monitoring.

In summary, the shift from categorical to dimensional emotional values in SER, as demonstrated in this research, opens a new way in understanding human emotions and offers an enhanced framework for developing more intuitive and emotionally aware human-computer interfaces.

ACKNOWLEDGMENT

This work was supported by the Industrial Technology Innovation Program (No. 20012603, Development of Emotional Cognitive and Sympathetic AI Service Technology for Remote Learning and Industrial Sites) funded By the Ministry of Trade, Industry and Energy (MOTIE, Korea).

REFERENCES

Atmaja, B. T., & Akagi, M. (2020). Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. APSIPA Transactions on Signal and Information Processing, 9, e17.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42, 335–359.

Elbarougy, R., & Akagi, M. (2014). Improving speech emotion dimensions estimation using a three-layer model of human perception. Acoustical Science and Technology, 35(2), 86–98.

Fang, X., Rychlowska, M., & Lange, J. (2022). Cross-cultural and inter-group research on emotion perception. Journal of Cultural Cognitive Science, 6(1), 1–7.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3451–3460.

Kim, J.-W., Kim, D.-H., Do, J.-S., & Jung, H.-Y. (2022). Strategies of utilizing pre-trained text and speech model-based feature representation for multi-modal emotion recognition. Proceedings of the Korean Information Science Society Conference, 2282–2284.

Konangi, U. M. Y., Katreddy, V. R., Rasula, S. K., Marisa, G., & Thakur, T. (2022). Emotion recognition through speech: A review. 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 1150–1153. https://ieeexplore.ieee.org/abstract/document/9792710/

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2020). Deep representation learning in speech processing: Challenges, recent advances, and future trends. ArXiv Preprint ArXiv:2001.00378. https://arxiv.org/abs/2001.00378

Letaifa, L. B., Torres, M. I., & Justo, R. (2020). Adding dimensional features for emotion recognition on speech. 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 1–6. https://ieeexplore.ieee.org/abstract/document/9231766/

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. L. (2023). Speech emotion recognition using machine learning—A systematic review. Intelligent Systems with Applications, 200266.

Noh, K. J., & Jeong, H. (2021). KEMDy19 [dataset]. https://nanum.etri.re.kr/share/kjnoh/KEMDy19?lang=ko_KR