Dement Neurocogn Disord. 2024 Jul;23(3):146-160
https://doi.org/10.12779/dnd.2024.23.3.146
pISSN 1738-1495·eISSN 2384-0757

DND Dementia and Neurocognitive Disorder

Original Article

Check for updates

Speech Emotion Recognition in People at High Risk of Dementia

Dongseon Kim,[1] **Bongwon Yi**,[2] **Yugwon Won**[3]

[1]Department of Silver Business, Sookmyung Women's University, Seoul, Korea
[2]Department of Communication Disorders, Korea Nazarene University, Cheonan, Korea
[3]Baikal AI Co. Ltd., Seoul, Korea

OPEN ACCESS

Received: May 10, 2024
Revised: Jul 8, 2024
Accepted: Jul 11, 2024
Published online: Jul 24, 2024

Correspondence to
Dongseon Kim
Department of Silver Business, Sookmyung Women's University, 100 Cheongpa-ro 47-gil, Yongsan-gu, Seoul 04310, Korea.
Email: weeny388@gmail.com

ORCID iDs
Dongseon Kim
https://orcid.org/0000-0003-1252-7873
Bongwon Yi
https://orcid.org/0000-0003-1312-8639
Yugwon Won
https://orcid.org/0000-0001-9419-9986

Funding
This work was supported by a National Research Foundation of Korea grant (2022R1I1A306104911 to DK).

Conflict of Interest
The authors have no financial conflicts of interest.

Author Contributions
Conceptualization: Kim D; Data curation: Kim D; Formal analysis: Won Y; Funding acquisition:

ABSTRACT

Background and Purpose: The emotions of people at various stages of dementia need to be effectively utilized for prevention, early intervention, and care planning. With technology available for understanding and addressing the emotional needs of people, this study aims to develop speech emotion recognition (SER) technology to classify emotions for people at high risk of dementia.
Methods: Speech samples from people at high risk of dementia were categorized into distinct emotions via human auditory assessment, the outcomes of which were annotated for guided deep-learning method. The architecture incorporated convolutional neural network, long short-term memory, attention layers, and Wav2Vec2, a novel feature extractor to develop automated speech-emotion recognition.
Results: Twenty-seven kinds of Emotions were found in the speech of the participants. These emotions were grouped into 6 detailed emotions: happiness, interest, sadness, frustration, anger, and neutrality, and further into 3 basic emotions: positive, negative, and neutral. To improve algorithmic performance, multiple learning approaches were applied using different data sources—voice and text—and varying the number of emotions. Ultimately, a 2-stage algorithm—initial text-based classification followed by voice-based analysis—achieved the highest accuracy, reaching 70%.
Conclusions: The diverse emotions identified in this study were attributed to the characteristics of the participants and the method of data collection. The speech of people at high risk of dementia to companion robots also explains the relatively low performance of the SER algorithm. Accordingly, this study suggests the systematic and comprehensive construction of a dataset from people with dementia.

Keywords: People at High Risk of Dementia; Speech Emotion Recognition; CNN+LSTM Algorithm; Deep Learning; Voice and Text Analysis

INTRODUCTION

Dementia is a debilitating syndrome, with enormous impact on people and societies. As the number of people at risk of dementia increases, the importance of preventing dementia in people at high-risk is being emphasized. Although there is limited evidence to prove a direct

Kim D; Investigation: Kim D; Methodology: Kim D, Yi B, Won Y; Project administration: Kim D; Software: Yi B; Supervision: Kim D; Validation: Kim D; Writing - original draft: Kim D; Writing - review & editing: Kim D, Yi B.

cause-and-effect relationship between preventive strategies and dementia, current studies suggest that a multifactorial approach may be the most promising to prevent cognitive decline. This approach includes regular exercise, a healthy diet, and the management of vascular diseases, psychosocial stress, and depression.[1] In particular, depressive disorder and psychological stress are strongly associated with increased risk of cognitive decline.[2-4] Depression and negative mood may further accelerate the progression of MCI to dementia.[5,6] Weng et al.[7] found insufficient social emotional support (SES) to be significantly associated with subjective cognitive decline (SCD). In this context, recognizing emotions is crucial for people at high risk, and remains essential as dementia progresses. Addressing both the physical and emotional needs of people with dementia is key to maintaining their quality of life. However, emotion recognition and support are insufficient for people who are at high risk, but have not yet been diagnosed for dementia. Emotional support is often lacking, even for people with dementia. Most dementia care focuses primarily on physical assistance, hygiene, and safety, often overlooking emotional needs, due to a shortage of caregivers. As recent technological interventions evolve, they are being considered a potential substitute for human caregiving. The related technology, in particular, emotion recognition through speech, is anticipated for both prevention and dementia care for people at various stages of dementia. Emotion recognition is further considered important in developing artificial intelligence (AI) and robot-based intervention, as it could serve as a facilitator to improve human–machine interaction. Although some robots already recognize basic human emotions, emotion recognition of people with dementia is still nascent.

Among various physical signs, such as facial expressions, gestures, and skin color, to recognize emotions,[8,9] this study focuses on speech as a medium for emotional expression. Speech, comprising voice tone, pitch, and content, is more obvious and externally accessible than other physical reflections of emotion.[10,11] Emotions are difficult to conceal in the voice due to the involuntary responses of vocal cords, governed by the sympathetic nervous system, supporting the necessity for voice-based emotion recognition.[12] Several studies have already explored voice features to identify stress and emotions in speech and vocalizations with statistical models, such as Hidden Markov model (HMM) and Gaussian mixture model (GMM).[13,14] According to investigations,[15,16] pitch, shimmer, Harmonic to Noise rate, Mel frequency cepstral coefficient, and the linear prediction cepstral coefficient are recognized as important features in emotions. Other than voice features, spoken content or text information is separately important to emotion recognition. Textual information retrieved from many sources, such as books, newspapers, web pages, and e-mail messages, is also rich in emotion. With the help of natural language processing techniques, emotions can be extracted from textual input by analyzing punctuation, emotional keywords, syntactic structure, and semantic information. Accordingly, attempts to recognize speech based solely on text, or by integrating voice and text, deserve attention.[17,18]

Despite being a longstanding area of study, the prior efforts to extract sound and text characteristics for speech emotion recognition (SER) were not completely successful. Though HMMs and GMMs excel at modeling time-series data and offer a relatively simple and interpretable framework for modeling sequences, they have limited capacity to model complex patterns in high-dimensional data, such as the nuanced emotions in speech.[19]

Recently, this area of study has regained attention due to the growing need for human–robot interfaces. Additionally, a relatively new approach utilizing deep learning methods, particularly convolutional neural network (CNN), deep CNN, and recurrent neural network

(RNN), is bringing forth advancements in this field. As deep learning gains momentum in the healthcare sector, research on SER is actively being conducted. These models showcase higher performance due to advances in technology. However, the application of SER using the speech of people with dementia remains largely experimental, primarily because such people are underrepresented in social activities.

This study seeks to analyze the emotional expression of people at high risk of dementia, and to develop a SER model employing deep neural learning. Motivated by the challenges in current technological development and the prospective advantages of employing advanced technologies in the prevention and care of dementia, this research explores the technological processes and discusses the implications of this technology for people at high risk of dementia.

METHODS

Data sources

Data for this study were collected using companion robots that were distributed to 133 community-dwelling older adults. All participants lived alone, and were included in the data collection because they had general cognitive impairments, such as depression and dementia in some cases. The data was not gathered directly by the researchers, nor were participants' cognitive abilities examined by the research team. Nevertheless, the participants were classified as being at high risk of dementia by the research team, due to their alleged cognitive impairment, social isolation, and age. Despite the heterogenous characteristics of participants considered, their state of depression and social isolation define them at high risk of dementia, as depression and social isolation are well-established risk factors of dementia.[20-22] In addition, physical and psychological health, cognition, and age are factors that strongly affect dementia.[23,24] The information provided to the research team consisted solely of gender and age. Among the participants, 104 were female (78.2%) and 29 were male (21.8%), with an average age of 80.6 years (standard deviation: 17.8). The age range varied 65 to 96 years of age. Advanced age is also a risk factor for dementia.[25]

Companion robots provided to participants were designed with a child-like appearance and programmed to ask questions, though unable to engage in natural conversation beyond that. Upon the robot's questions, participants spoke on what had happened during the day, and how they felt on a daily basis. The participants treated the robots as if they were their own grandchildren, while also acknowledging their non-human nature. They expressed appreciation, stating phrases like "you are good at talking, even without a mouth." The data collection continued for three months. The monologues produced by participants were sent to a server for researchers to access and download. This voice data amassed around 11,000 audio files, ranging from 30 seconds to 2 minutes for each file, cumulatively amounting to 13.1 GB in digital size. A meticulous selection process, based on audio quality, resulted in the retention of 6,899 recordings (2.27 GB). Recordings marred by background noise, such as from television, radio, and other external sources, were excluded from the analysis.

Data labelling: emotion classification by human auditory assessment

Machine learning approaches can be classified into two types based on data preprocessing and handling: unsupervised, where training involves raw data and automatic feature extraction, and supervised, which involves manual feature extraction to aid classification. Comparing these systems is challenging due to various factors, such as the dataset, classified

https://doi.org/10.12779/dnd.2024.23.3.146

categories, neural network hyperparameters, training procedures, and evaluation metrics. The decision to use either supervised or unsupervised methods is at the discretion of the researcher.[26] This study employed a supervised approach, incorporating data labeling or annotation based on human auditory assessment for emotion classification. Three trained graduate students were responsible for transcribing the audio data and classifying emotions. The three coders analyzed 6,899 voice files through a sequential auditory classification process. They initially categorized a broad spectrum of emotions without predefined categories, progressively narrowing them down to 6 specific emotions, and ultimately to 3 categories of negative, neutral, and positive. Categorization primarily relied on voice tone and pitch. Since voice tone and content are inseparable in speech analysis, both are included in the assessment. However, voice tone takes precedence, as exemplified when a participant says "I feel better" in a subdued voice. Emotions were named by referencing Plutchik's Wheel. Plutchik's wheel categorizes emotions into 8 primary categories of joy, trust, fear, surprise, sadness, anticipation, anger, and disgust, along with their respective subcategories and potential combinations.[27] During the classification stages, the coders participated in training workshops to improve emotional awareness, and ensure consistency in classifications. They were trained to listen to the same audio files, and discuss their classifications to promote consistency. All three graduate students were given the same audio files, and their agreement (inter-rater reliability) was assessed using the Kappa statistic.

Deep learning architecture: CNN + long short-term memory (LSTM) + Attention layers

The design of the architecture–detailing layer depth, width, and types–usually affects the model's learning ability, performance, and problem-solving efficiency. For high performance, this model incorporated Wav2Vec 2.0 for the novel speech extraction features, LSTM layers, and attention mechanism in the CNN-based architecture. These elements work together to extract voice features, understand long-term data patterns, focus on important information, and accurately predict outcomes. Text was also included in the training. For text-based learning, KPFBERT, a variation of BERT30, was utilized to create a learning model that captures the nuances of the emotions expressed in each text.

Below are descriptions of the key processes involved in the model. Firstly, the modeling commences with the extraction of robust features from audio data utilizing the Wav2Vec 2.0 framework, which was developed by Meta AI. Wav2vec 2.0 represents a significant advancement in deep learning techniques for speech analysis. It captures the nuanced acoustic properties of speech by pre-training on a vast corpus of unlabeled audio data, which yields high-quality feature extraction.[28] Secondly, the features extracted by Wav2Vec 2.0 are input to a CNN–LSTM model designed to leverage both the voice feature extraction capabilities of CNN,[29] and the sequential data processing strengths of LSTM networks.[30] The CNN component comprises multiple convolutional layers with varying kernel (dimensions of the filter) sizes, capturing a broad range of acoustic patterns across different scales. Subsequently, the LSTM layers analyze how speech features change over time, effectively grasping the long-lasting connections crucial for comprehending the emotional tone of speech, resulting in improved quality of SER.

Thirdly, an attention mechanism was integrated into LSTM processing to enhance the model's focus on the most informative parts of the speech, and offer a more nuanced understanding of the emotional signals in speech.[31] Recent studies, including research by Kumar et al.,[32] have already demonstrated the effectiveness of CNN–LSTM models in speech

https://doi.org/10.12779/dnd.2024.23.3.146

classification. By incorporating the Wav2Vec 2.0 for enhanced extraction capabilities and attention mechanism for the assignment of weighted values, our model aims to set a new benchmark for accuracy and efficiency in speech emotion recognition.

In this study, we also developed a novel text-based model with a pretrained language model KPFBERT, a variation of BERT. The process starts with tokenization and attention masks for the extraction of emotional nuances from text. Tokenization is the process of breaking down sentences into words or smaller units, while attention masks help the model focus on important words, and distinguish the roles of words within a sentence. The extracted information is passed through 2 fully connected layers with ReLU activation functions to classify emotions. The ReLU activation function applies a non-linear transformation to the outputs, enhancing the model's ability to represent complex patterns. Then, CNN + LSTM layers and attention mechanisms were also employed to train the extracted text-based emotional features. **Fig. 1** shows the architecture of the CNN + LSTM + Attention model.[28]

The data training process

The training process consists of training, validation, and testing. Thus, the entire dataset is divided into 'Training,' 'Validation,' and 'Testing' subsets for the respective purposes of training, tuning, and evaluating the model. Notably, validation and testing are often misconstrued concepts, yet they possess distinct differences. Validation data evaluates and adjusts hyperparameters during the model's training phase, preventing overfitting, and assessing the model's generalizability to new data. Overfitting refers to excessive learning, where a model becomes overly tailored to the training data, to the extent that the model may fail to predict on new data. Avoiding overfitting involves achieving a balance in learning, determined by loss and accuracy values, which vary with the number of epochs. An epoch refers to the process of a model learning from a training dataset. To enhance the model's performance, it is necessary to train data by repeating epochs, until the loss reaches its minimum, and the accuracy reaches its maximum. 'Loss' represents the error of the model on the validation or testing data. The lower the loss value, the better the performance of the model is considered. 'Accuracy' is the ratio that indicates how accurately the model has predicted in classification. For example, if it correctly predicted 90 out of 100 samples, then the accuracy is 90%.

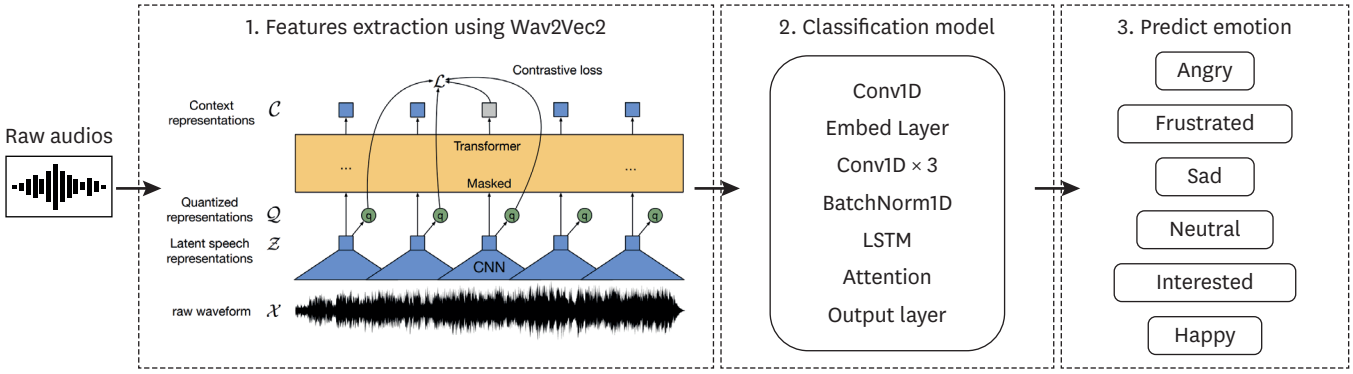

Fig. 1. The architecture of the CNN + LSTM + Attention model.
Drawn by the authors, based on Baevski et al. (2020).[28]
CNN: convolutional neural network, LSTM: long short-term memory.

https://doi.org/10.12779/dnd.2024.23.3.146

The performance of the model also varies with the number of emotion categories and the inclusion of text content. Therefore, the authors experimented with classifying emotions into three simple categories of (good, bad, and neutral), or more detailed subdivisions.

To evaluate the performance of the voice-based and text-based algorithm, precision, recall, and the F1-score of each model were referred.

Ethical issues

Data collection was carried out by a separate research team, with the authors not directly involved in the data collecting process. The authors could not access any information other than the gender and age of the participants, which did not allow the authors to identify or contact them to gain their consent. Due to the circumstances, the Institutional Review Board of the Sookmyung Women's University granted this study an exemption from ethical review (No. 1041549-230411-SB-166). Furthermore, the authors faced no issues related to participant privacy during the audio file analysis, as the recordings contained no personal information, such as names, addresses, or banking details. Nevertheless, precautions were implemented to safeguard the information, including its storage on a specified medium, and restricting the research group's access solely for research-related objectives.

RESULTS

Emotional classification by human auditory assessment

The notable aspect highlighted by this study on emotion classification in people at high risk of dementia is to inspect the breadth of their emotions, and determine how their expressed emotions differ from those of the general population. Emotion classification in this study was hierarchically conducted in several stages to enhance classification accuracy. During the initial phase, a total of 27 unique emotions, such as excitement, willingness, boasting, praise, complaint, tiredness, and others, were discerned through human auditory assessment, which involved collaborative decision-making through discussions in cases of disagreement. Inter-coder agreement during this phase, assessed with SPSS's reliability K-value, demonstrated a moderate level of concordance at 0.70. The second phase was to categorize the 27 emotions into simpler groups: happy, interested, angry, sad, frustrated, and neutral were included. 'Interested' posed the greatest challenge in naming. As it denotes a positive emotion, albeit with subdued energy compared with happiness, there was a need to categorize such emotions into a separate group. For example, 'Interested' was named in accord with Plutchik's 'trust' and 'anticipation.' Finally, positive, negative, and neutral were grouped. From a quantitative perspective, in terms of the number of data, 'Neutral' (n=2,634, 38.2%) was the most frequent, followed by 'interest' (n=2,076, 30.1%), 'happiness' (n=754, 10.9%), 'frustration' (n=638, 9.3%), 'sadness' (n=582, 8.45%), and 'anger' (n=215, 3.1%) (**Fig. 2**).

Performance of the CNN-based model for SER

Speech from participants was classified by human auditory assessment into six emotions, and then put into CNN + LSTM architecture for feature extraction and training. The performance of the model was evaluated with the precision, recall, and F1-score for each emotion. Precision means the ratio of cases where the model predicts 'positive,' and they are actually 'positive.' For example, the model with high precision value accurately identifies an 'interested' emotion when it is indeed 'interested.' Recall refers to the proportion of 'angry' cases detected among all the 'angry' cases, its detectability without missing any. F1-score is the harmonic mean of the precision and recall.

https://doi.org/10.12779/dnd.2024.23.3.146

Fig. 2. Hierarchical classification of emotions by human auditory assessment.

$$F1 = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

The F1-score is useful for evaluating the performance of a balanced model, ensuring it does not favor one side over the other. For example, if precision is high but recall is low, or *vice versa*, the F1-score decreases. With precision, recall and F1-score, the performance of the model was reexamined with loss and accuracy in the stage of validation and testing.

Initially, training was performed using voice data labelled with six different emotions. The results were examined by emotion (**Table 1**), with 'happy' having the highest precision (0.6429), followed by 'interested' (0.4906), 'sad' (0.4733), 'frustrated' (0.4638), and 'angry' (0.4561). Conversely, in terms of recall, 'sad' (0.625) was highest, followed by 'angry' (0.5778), 'frustrated' (0.4571), 'happy' (0.5), and 'interested' (0.3611). As the values of precision and recall were low, the performances in both validation and testing were also unsatisfactory. The validation accuracy stood at 0.4836, while the test accuracy reached 0.5141.

https://doi.org/10.12779/dnd.2024.23.3.146

Table 1. Detection rates across models varying in the number of emotions and data types

Resources	Label	Precision	Recall	F1-score	Data count
Voice-based	Angry	0.4561	0.5778	0.5098	440
	Frustrated	0.4638	0.4571	0.4604	720
	Sad	0.4737	0.625	0.5389	720
	Neutral	0.4247	0.4306	0.4276	720
	Interested	0.4906	0.3611	0.416	720
	Happy	0.6429	0.5	0.5625	720
	Average	0.4919	0.4919	0.4858	720
Voice-based	Negative	0.6628	0.6096	0.6351	187
	Positive	0.6595	0.6421	0.6507	190
	Neutral	0.6273	0.69	0.6571	200
Text-based	Negative	0.8227	0.8171	0.8354	327
	positive	0.9167	0.8486	0.864	433
	Neutral	0.8079	0.8746	0.8399	187

As the overall results on the six emotions based on voice features were unsatisfactory, it was necessary to explore new learning methods. As the performance of the model varies depending on the number of emotions to be classified, the next training was performed with the data of 3 basic emotions, of good, bad, and neutral. To classify the three basic emotions, the model was trained in 2 ways, each based on voice and text. Speech encompasses not only voice, but also content. Excluding the tone of voice, text can sometimes be ambiguous; however, the content, being either positive or negative, needs to be included to determine emotions.[17] Upon comparing the results, it was found that text-based classification outperformed voice-based classification by a significant margin (**Table 1**). When classifying the 3 emotions through text, precision and recall of 91.67% and 84.86%, respectively, for positive (F1-score=0.86), and 82.27% and 81.71%, respectively, for negative (F1-score=0.84) were achieved. It was also observed that positive emotions were classified more accurately in the text-based emotion classification.

Concerning the voice-based classification, the precision for identifying negative emotions was 66.3%, while for positive emotions it was 66.0%, showing no significant difference between the 2. The recall rate (64.2%) for positive emotions was slightly higher than that of negative emotions (61.0%). As result, the accuracies in detecting the three basic emotions also rose to 0.6956 for validation, and 0.6474 for test (average=0.6715).

After dividing emotions into 3 categories of positive, negative, and neutral, the model was trained to further classify into detailed emotions within each category. This stepwise approach was utilized with voice-only, and multimodal of text and voice. The voice-only 2-stage classification model results in an increase of precision and recall from the previous models. For 'interested,' the 2-step approach achieved precision of 73.7% (recall=58.3%, F1=0.6512) which is significantly higher than the 49.06% for the first classification model based on voice only (**Table 1**). Similarly, the precision for 'angry' improved from 45.61% to 57.7% (recall=52.9%, F1=0.5521), and for 'sad' from 47.37% to 61.0% (recall=68.0%, F1=0.6434), indicating an overall enhancement in precision. It is worth interpreting the precision and recall for each emotion. For example, a precision of 73.7% and a recall of 58.3% for 'interested' means that the emotion 'interested' has a high probability of being correctly identified, but it is less frequently detected among all instances of the 'interested' emotion. Compared to positive emotions, negative emotions were less detected. 'Sad' was detected with precision of 61.0% (recall=68.0%, F1=0.6434), followed by 'angry' and 'frustrated.' 'Angry' and 'frustrated' were still difficult for the model to classify. Notwithstanding, the performance of the two-step recognition process showed improvement over that of one-step classification.

https://doi.org/10.12779/dnd.2024.23.3.146

Table 2. Detection rate in the voice and text-based classifications in the 2-stage approach

Stages of classification	Label	Precision	Recall	F1-score	Data count
Voice-based two-stage classification					
Negative	Sad	0.6103	0.6803	0.6434	122
	Angry	0.5769	0.5294	0.5521	85
	Frustrated	0.5752	0.5417	0.5579	120
Positive	Interested	0.7368	0.5833	0.6512	216
	Happy	0.6565	0.7926	0.7182	217
Neutral		0.6423	0.5577	0.5656	187
Combined with Text and Voice + two-stage classification					
Negative	Angry	0.7108	0.6555	0.682	85
	Sad	0.7284	0.7586	0.7432	122
	Frustrated	0.6614	0.6086	0.6339	120
Positive	Interested	0.8439	0.6728	0.7487	216
	Happy	0.5879	0.8106	0.6815	217
Neutral		0.7065	0.7012	0.6979	187

Finally, a 2-stage and multimodal approach was utilized, involving text to categorize the 3 basic emotions, followed by voice-based classification for more nuanced emotions. In terms of precision, interested (84.4%) was the highest, followed by sad, angry, frustrated, and happy at (72.8%, 71.1%, 66.1%, and 58.8%, respectively. Recall shows that happy was the highest (81.1%), followed by sad, interested, angry, and frustrated at 75.9%, 67.3%, 65.6%, and 60.9%, respectively. Happy emotions are the most distinctly expressed, yet identifying them accurately is challenging. Frustration is difficult to both predict and detect. However, the accuracy of each emotion is difficult to rank, as accuracy requires consideration of precision, recall, and overall accuracies in the validation and test processes. These results can be compared with the results of the voice-only learning model (**Table 2**). Based on the comparison, the 2-staged and multimodal model demonstrated overall better results, compared to a model conducted in two stages using only voice.

Table 3 shows the validation and testing results of the various models explored in this study. The initial six emotions classification model based on voice yielded accuracy of 48.4% for validation, and 51.4% for testing. Voice-based recognition of the three emotions yielded scores of 69.6% and 64.7% for validation and testing, respectively, while text-based recognition on basic emotions showed higher scores of 83.4% and 79.8%, respectively. The multimodal two-step model, which classifies 6 emotions, achieved accuracies in validation and testing of 70.1% and 69.3%, respectively. This means that the model shows the highest performance in predicting detailed emotions. **Fig. 3** better explains the meanings of loss and accuracy in validation and testing. In the right and middle graphs of **Fig. 3**, as the number of epochs progresses, 'train loss' consistently decreases, and 'train accuracy' increases. However, in the same graphs, there is no more discernible increase of accuracy past a certain number of epochs, which suggests saturation in the model's ability to predict new data accurately. Examining these graphs reveals that conducting training for approximately 4–5

Table 3. Performance of different training models

Outcomes	Process	Loss	Accuracy
Six emotions	Validation	1.344	0.4836
	Test	1.29	0.5141
Three emotions	Validation	0.6803	0.6959
	Test	4.141	0.6474
Text-based three emotions	Validation	5.034	0.8344
	Test	5.852	0.7984
Text and voice-based six emotions	Validation	9.262	0.7009
	Test	9.706	0.6934

https://doi.org/10.12779/dnd.2024.23.3.146

DND Dementia and Neurocognitive Disorder

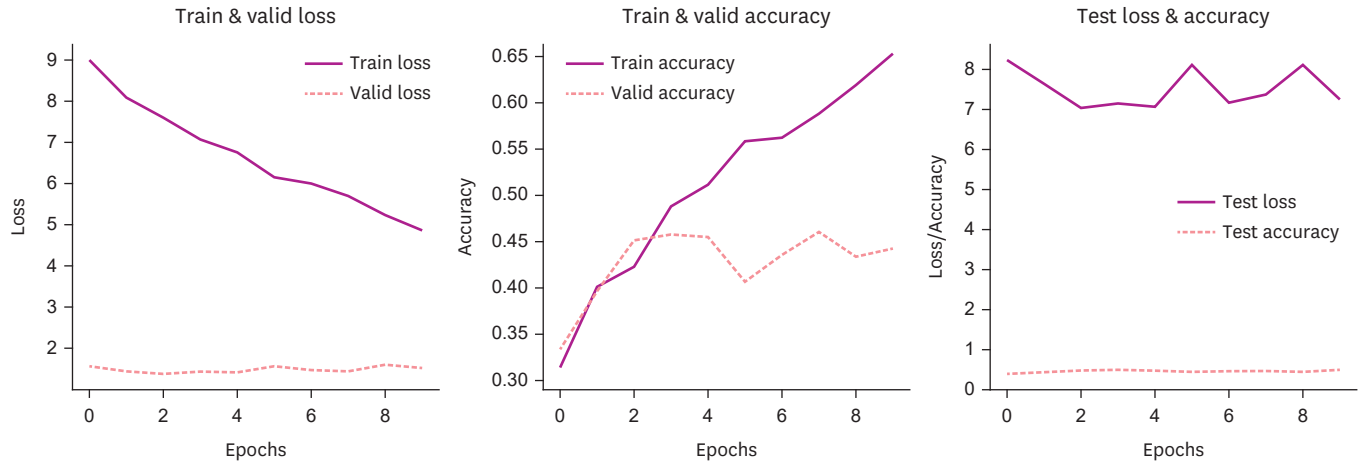

Fig. 3. Loss and accuracy graphs of training, validation, and testing on six emotions.

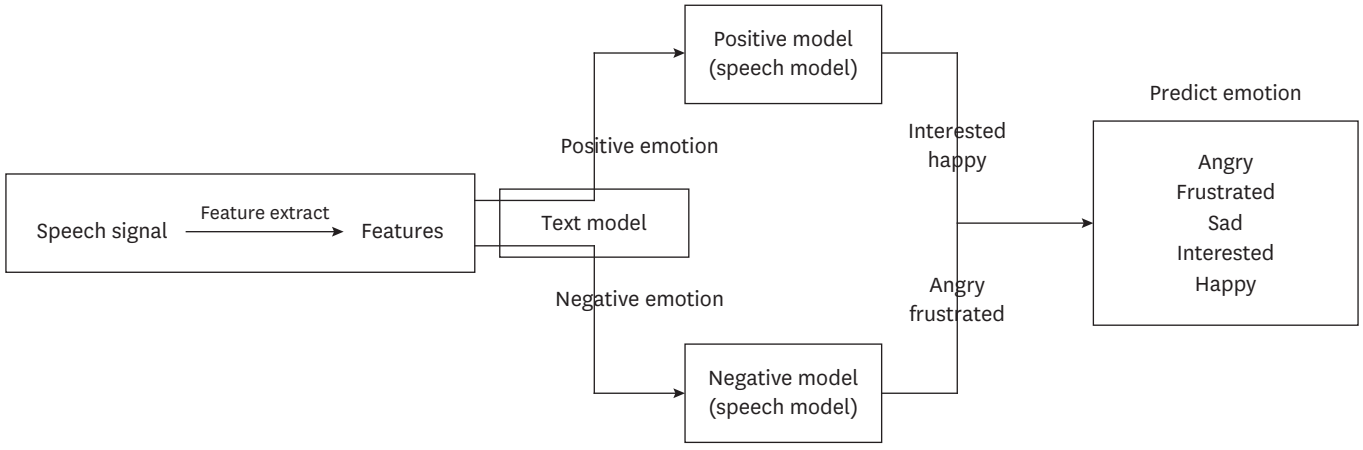

Fig. 4. Model flowchart of the hierarchical model.

epochs would be prudent, as further training beyond this point does not yield any additional performance improvements.

After conducting a series of trials and analyzing the results, this study selected the final model that exhibited the highest performance (**Fig. 4**).

DISCUSSION

Recognizing the emotions of people at the various stages of dementia is recommended for prevention, early intervention, and care planning. Emotion recognition is becoming increasingly essential as technology advances in dementia care, complementing, or even replacing, human care. Efforts to recognize emotions from speech have progressed from earlier approaches that extracted vocal characteristics with statistical models to the current utilization of deep learning methods. This research analyzed the emotions and developed the SER model with the speech of people at high risk of dementia. The process involved emotion classification by human auditory assessment, and building various deep learning models to enhance the performance of the model. Here, the authors aim to discuss the key findings and

https://doi.org/10.12779/dnd.2024.23.3.146

limitations obtained through the research procedures. Firstly, this study could find diverse and vivid emotions from the voices of the participants. Human coders identified 27 distinct emotions, which were then categorized into smaller groups for subsequent deep learning modelling. These emotions ranged from positives of excitement, boasting, and gratitude, to negatives of complaint, boredom, and anxiety. Also, note that positive emotions (41.0%) were more prevalent than negative emotions (20.85%). The results found in this study can be attributed to the heterogeneous characteristics of the participants, rather than their grouping as dementia patients. In general, the prevailing stereotypical portrayals depict people with dementia as constantly in a diminished emotional state, depressed, and exhibiting anxiety and agitation.[33-36] Bucks and Radford[33] reported that patients of Alzheimer's disease have a deficit of emotional processing, compared to healthy elderly adults. Han et al.[37] showed that people in the early stage of dementia have impaired emotion expression on the retrieval of autobiographical memories. According to the related studies, these deficits are secondary to the primary cognitive impairments of the patients.[38,39] In addition, this stereotype has been enforced as people with dementia experience disease-related language, comprehension, and memory deficits that preclude the self-report of emotions.[40]

Apart from the literature review, the positive and diverse emotions found in this study might be a reflection of the participants being at the pre-onset, or early stage, of the disease, leading to the authors speculating that these people can maintain a rich inner state through appropriate prevention and intervention. Additionally, the method of data collection likely contributed to the diversity and numerical imbalance of emotions, with neutral emotions accounting for 38.2%. The data were gathered using companion robots that served as intermediaries for the speech of participants. Generally, the main method of collecting emotion-related speech involves simulating emotional expression, namely, asking actors to produce vocal expressions of specific emotions.[41] The popular datasets for SER, the Crowd-sourced Emotional Multimodal Actors Dataset, Ryerson Audio–Visual Database of Emotional Speech and Song, Surrey Audio–Visual Expressed Emotion, Toronto Emotional Speech Set, and the Interactive Emotional Dyadic Motion Capture were built with recordings of actors under a laboratory environment.[42,43] The datasets were constructed specifically for the purpose of emotion classification, ensuring a balanced distribution of all emotion categories by design. In contrast, the dataset for this study was made up of older people living alone in natural situations, which contributed to a numerical imbalance of emotions. It is worth mentioning that even deliberately constructed datasets sometimes exhibit an imbalance in the proportion of emotions. The National Information Society Agency's 'Emotion-Tagged Free Conversation' corpus, which was constructed with a huge budget and is publicly available for Korean sentiment classification, also reveals an imbalance of emotions. The emotional proportion of this corpus is as follows: happy (38.13%), surprised (13.14%), afraid (2.78%), loving (8.66%), sad (6.61%), angry (7.8%), and unknown (22.88%). Classified by emotional type, positive emotions (69.17%) outnumbered negative emotions (18.34%), with 12.48% of neural emotion.[44]

The vivid and genuine emotional expression in the dataset of this study is also attributed to data collection. Induced by companion robots, the speech was not controlled acting, but spontaneous and genuine. The participants may have been freed of social constraints while engaging with companion robots, without concerns about being overheard by others. They were heard to frankly vocalize their emotions by sometimes exploding in anger, or singing out of joy, which they would not normally do in front of other people.

https://doi.org/10.12779/dnd.2024.23.3.146

The second finding is also related to the dataset, which concerns the accuracy rate of the CNN-based model for SER. The model in this study underwent a process of performance enhancement, resulting in 70% with a two-staged approach: a preliminary text-based categorization, followed by detailed emotion analysis based on voice. Performance of SER modelling usually depends on good architecture and a reliable dataset. Regarding architecture, this model is based on CNN + LSTM, the architecture that is currently commonly utilized, and integrated with attention layers, which are quite highly functional. Additionally, this study's model has a modern architecture using Wav2Vec 2.0 for effective voice feature extraction and attention mechanism, which focuses on key parts of a speech sequence to weight each segment, and better understand emotions. Considering the algorithm's innovative potential presented in other studies, the final accuracy rates are deemed insufficient. The recent application of CNN and LSTM has achieved notable success in other research efforts, with benchmarks reporting accuracy that ranges 80% to 95%.[45,46] In this regard, the low performance needs to be investigated further with the characteristics of dataset. As previously mentioned, models demonstrating high performance typically use datasets that consist of high-quality audio and balanced data collected in controlled laboratory environments. In contrast, this SER model was built and evaluated using heterogeneous data from people at high risk of dementia under the natural environment. As the audio data were recorded in the participant's residence, natural background noises, excessively short phonations, and other sounds were included, compromising the quality. The uncontrolled and genuine nature of our data reveals potential issues for SER applications in real-world settings, demonstrating the challenges of algorithms developed in laboratories. Thus, the low performance observed in this study should not be seen merely as a limitation of the model, but rather as a necessary step toward understanding the challenges of building models in specialized domains.

Unlike traditional statistical approaches, deep learning requires extensive and well-constructed training data. In this regard, the authors propose the development of a comprehensive and systematic voice dataset for people with dementia. The pathological features of speech in people with dementia have been currently used for diagnosis,[47,48] indicating that their voice data is likely different from that of the general population. The limitations in using datasets created from the performances of actors seem apparent in learning about the everyday language and emotional expressions of people with dementia.

Considering the typically limited representation of people with dementia in both academic and social contexts, their diverse representation (age, gender, education, dialects, the stage of dementia progression, etc.) needs to be included for the quality of the dataset.

Another suggestion from this study involves actively introducing SER to leverage technological achievements in prevention and dementia care. Since emotions are indicators of quality of life, identifying them plays a vital role in prevention, intervention, and providing appropriate care. This study highlights the potential of using rapidly advancing technology for prevention and emotional support for people at various stages of dementia as a key future endeavor.

https://doi.org/10.12779/dnd.2024.23.3.146

ACKNOWLEDGEMENTS

This research was made possible, in part, using the data provided by Hyodol Co. Ltd., a manufacturer of companion robots in Korea. The data collected through the companion robot, Hyodol, was provided to the researcher team free of charge for the purpose of the study, without any other conditions or duties.

REFERENCES

1. Sutin AR, Stephan Y, Terracciano A. Psychological well-being and risk of dementia. Int J Geriatr Psychiatry 2018;33:743-747. PUBMED | CROSSREF
2. Katon W, Pedersen HS, Ribe AR, Fenger-Grøn M, Davydow D, Waldorff FB, et al. Effect of depression and diabetes mellitus on the risk for dementia: a national population-based cohort study. JAMA Psychiatry 2015;72:612-619. PUBMED | CROSSREF
3. Ownby RL, Crocco E, Acevedo A, John V, Loewenstein D. Depression and risk for Alzheimer disease: systematic review, meta-analysis, and metaregression analysis. Arch Gen Psychiatry 2006;63:530-538. PUBMED | CROSSREF
4. da Silva J, Gonçalves-Pereira M, Xavier M, Mukaetova-Ladinska EB. Affective disorders and risk of developing dementia: systematic review. Br J Psychiatry 2013;202:177-186. PUBMED | CROSSREF
5. Richard E, Reitz C, Honig LH, Schupf N, Tang MX, Manly JJ, et al. Late-life depression, mild cognitive impairment, and dementia. JAMA Neurol 2013;70:374-382. PUBMED | CROSSREF
6. Mourao RJ, Mansur G, Malloy-Diniz LF, Castro Costa E, Diniz BS. Depressive symptoms increase the risk of progression to dementia in subjects with mild cognitive impairment: systematic review and meta-analysis. Int J Geriatr Psychiatry 2016;31:905-911. PUBMED | CROSSREF
7. Weng X, George DR, Jiang B, Wang L. Association between subjective cognitive decline and social and emotional support in US adults. Am J Alzheimers Dis Other Demen 2020;35:1533317520922392. PUBMED | CROSSREF
8. Lawton MP, Van Haitsma K, Klapper J. Observed affect in nursing home residents with Alzheimer's disease. J Gerontol B Psychol Sci Soc Sci 1996;51:3-14. PUBMED | CROSSREF
9. Vogelpohl TS, Beck CK. Affective responses to behavioral interventions. Semin Clin Neuropsychiatry 1997;2:102-112. PUBMED | CROSSREF
10. Ekman P. *Emotions Revealed, Second Edition: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. New York: Henry Holt and Company, 2007.
11. Izdebski K. *Emotions in the Human Voice. Volume 1, Foundations*. San Diego: Plural Publishing, Inc., 2008.
12. Higuchi M, Nakamura M, Shinohara S, Omiya Y, Takano T, Mitsuyoshi S, et al. Effectiveness of a voice-based mental health evaluation system for mobile devices: prospective study. JMIR Form Res 2020;4:e16455. PUBMED | CROSSREF
13. Kwon OW, Chan K, Hao J, Lee TW. Emotion recognition by speech signals. In: Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003); 2003 Sep 1-4; Geneva, Switzerland. International Speech Communication Association, 2003; 125-128.
14. Nogueiras A, Moreno A, Bonafonte A, Mariño JB. Speech emotion recognition using hidden Markov models. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001); 2001 Sep 3-7; Aalborg, Denmark. International Speech Communication Association, 2001; 2679-2682.
15. Hansen JH. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Commun 1996;20:151-173. CROSSREF
16. Bou-Ghazale SE, Hansen JH. A comparative study of traditional and newly proposed features for recognition of speech under stress. IEEE Trans Speech Audio Process 2000;8:429-442. CROSSREF
17. Chuang ZJ, Wu CH. Multi-modal emotion recognition from speech and text. Int J Comput Linguist Chin Lang Process 2004;9:45-62.
18. Yoon S, Byun S, Jung K. Multimodal speech emotion recognition using audio and text. arXiv. Forthcoming 2018.
19. Lu Q, Sun X, Long Y, Gao Z, Feng J, Sun T. Sentiment analysis: comprehensive reviews, recent advances, and open challenges. IEEE Trans Neural Netw Learn Syst 2023;PP:1-21. PUBMED | CROSSREF

20. Livingston G, Huntley J, Sommerlad A, Ames D, Ballard C, Banerjee S, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. Lancet 2020;396:413-446. PUBMED | CROSSREF
21. National Institute for Health and Care Excellence (NICE). *Dementia, Disability and Frailty in Later Life – Mid-Life Approaches to Delay or Prevent Onset*. London: NICE, 2015.
22. Daviglus ML, Plassman BL, Pirzada A, Bell CC, Bowen PE, Burke JR, et al. Risk factors and preventive interventions for Alzheimer disease: state of the science. Arch Neurol 2011;68:1185-1190. PUBMED | CROSSREF
23. Saczynski JS, Beiser A, Seshadri S, Auerbach S, Wolf PA, Au R. Depressive symptoms and risk of dementia: the Framingham Heart Study. Neurology 2010;75:35-41. PUBMED | CROSSREF
24. Barnes DE, Alexopoulos GS, Lopez OL, Williamson JD, Yaffe K. Depressive symptoms, vascular disease, and mild cognitive impairment: findings from the Cardiovascular Health Study. Arch Gen Psychiatry 2006;63:273-279. PUBMED | CROSSREF
25. Power MC, Mormino E, Soldan A, James BD, Yu L, Armstrong NM, et al. Combined neuropathological pathways account for age-related risk of dementia. Ann Neurol 2018;84:10-22. PUBMED | CROSSREF
26. Escobar-Linero E, Luna-Perejón F, Muñoz-Saavedra L, Sevillano JL, Domínguez-Morales M. On the feature extraction process in machine learning. An experimental study about guided versus non-guided process in falling detection systems. Eng Appl Artif Intell 2022;114:105170. CROSSREF
27. Mondal A, Gokhale SS. Mining emotions on Plutchik's wheel. In: Proceedings of the 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS); 2020 December 14-16; Paris, France. Piscataway; IEEE, 2020; 1-6.
28. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. Adv Neural Inf Process Syst 2020;33:12449-12460.
29. Lecun Y, Bottou Y, Bengio P, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86:2278-2324. CROSSREF
30. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9:1735-1780. PUBMED | CROSSREF
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst 2017:30.
32. Kumar Y, Koul A, Singla R, Ijaz MF. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput 2023;14:8459-8486. PUBMED | CROSSREF
33. Bucks RS, Radford SA. Emotion processing in Alzheimer's disease. Aging Ment Health 2004;8:222-232. PUBMED | CROSSREF
34. Mega MS, Cummings JL, Fiorello T, Gornbein J. The spectrum of behavioral changes in Alzheimer's disease. Neurology 1996;46:130-135. PUBMED | CROSSREF
35. Algase DL, Beck C, Kolanowski A, Whall A, Berent S, Richards K, et al. Need-driven dementia-compromised behavior: an alternative view of disruptive behavior. Am J Alzheimer Dis 1996;11:10-19. CROSSREF
36. Gonçalves-Pereira M. Neuropsychiatric symptoms in cognitive impairment and dementia: a brief introductory overview. In: Verdelho A, Gonçalves-Pereira M. *Neuropsychiatric Symptoms in Cognitive Impairment and Dementia*. Cham; Springer, 2017; 1-7.
37. Han KH, Zaytseva Y, Bao Y, Pöppel E, Chung SY, Kim JW, et al. Impairment of vocal expression of negative emotions in patients with Alzheimer's disease. Front Aging Neurosci 2014;6:101. PUBMED | CROSSREF
38. Cadieux NL, Greve KW. Emotion processing in Alzheimer's disease. J Int Neuropsychol Soc 1997;3:411-419. PUBMED | CROSSREF
39. Zandi T, Cooper M, Garrison L. Facial recognition: a cognitive study of elderly dementia patients and normal older adults. Int Psychogeriatr 1992;4:215-221. PUBMED | CROSSREF
40. Hahn EA. Daily experiences in stress, memory, and emotion in older adults with mild cognitive impairment [dissertation]. Tampa: University of South Florida, 2012.
41. Shah Fahad M, Ranjan A, Yadav J, Deepak A. A survey of speech emotion recognition in natural environment. Digit Signal Process 2021;110:102951. CROSSREF
42. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PLoS One 2018;13:e0196391. PUBMED | CROSSREF
43. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech 2005); 2005 Sep 4-8; Lisbon, Portugal. International Speech Communication Association, 2005; 1517-1520.

44. National Information Society Agency (NIA). Free conversation with emotion tags (adult) [Internet]. Daegu: NIA; 2022 [cited 2024 Jun 12]. Available from:https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=71631.
45. Mohamed O, Aly SA. Arabic speech emotion recognition employing Wav2vec2. 0 and HuBERT based on BAVED dataset. arXiv. Forthcoming 2021. CROSSREF
46. Al-onazi BB, Nauman MA, Jahangir R, Malik MM, Alkhammash EH, Elshewey AM. Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion. Applied Sciences. 2022;12:9188. CROSSREF
47. Xue C, Karjadi C, Paschalidis IC, Au R, Kolachalama VB. Detection of dementia on voice recordings using deep learning: a Framingham Heart Study. Alzheimers Res Ther 2021;13:146. PUBMED | CROSSREF
48. Park CY, Kim M, Shim Y, Ryoo N, Choi H, Jeong HT, et al. Harnessing the power of voice: a deep neural network model for Alzheimer's disease detection. Dement Neurocogn Disord 2024;23:1-10. PUBMED | CROSSREF