Multimodal Emotion Recognition and Dataset Construction in Online Counseling

Toshiki Takanabe, Kotaro Kashihara, Kazuyuki Matsumoto, Keita Kiuchi, Xin Kang, Ryota Nishimura, Manabu Sasayama

Proceedings of the 38th Pacific Asia Conference on
Language, Information and Computation (PACLIC 38)

Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, Jong-Bok Kim (eds.)

2024

Multimodal Emotion Recognition and Dataset Construction in Online Counseling

Toshiki Takanabe, Kotaro Kashihara, Kazuyuki Matsumoto, Keita Kiuchi,
Xin Kang, Ryota Nishimura, Manabu Sasayama
Tokushima University, 2-1 Minamijousanjima-cho, Tokushima-shi, Tokushima 770-0814, Japan
c612435041@tokushima-u.ac.jp, matumoto@is.tokushima-u.ac.jp

Abstract

In this study, we developed a multimodal dataset and performed emotion recognition experiments. The dataset includes objective emotion labels derived from online counseling videos. Five individuals were asked to predict the emotions of the person speaking in each counseling video and to assign emotion labels. Each video was evaluated by positioning a cursor on Russell's circumplex model, where the x-axis represents emotional valence (pleasantness-unpleasantness) and the y-axis represents arousal levels. To assess the inter-rater reliability of these evaluations, we calculated Fleiss' kappa. Using the constructed dataset, we conducted an emotion recognition experiment employing a Hybrid Fusion approach. Specifically, we used emotion recognition results from py-feat as features from images, acoustic features from wav2vec2.0 as features from speech and text-embedding-3 as features from language. When the acoustic features were weighted 0.4, the facial features 0.3, and the linguistic features 0.3, the result for the 16 emotion classifications was the most accurate, with a score of 0.4521.

1 Introduction

The COVID-19 pandemic has rapidly accelerated the adoption of online counseling. However, one of the challenges of online counseling is the difficulty in accurately identifying subtle facial expressions and vocal tones. To assist counselors in their assessments and improve operational efficiency, automatic emotion analysis of clients is considered to be highly effective. When humans interpret the emotions of others, they rely on a comprehensive judgment based on multiple cues, such as vocal tone, facial expressions, and speech content. Similarly, emotion estimation by AI can achieve high accuracy through multimodal emotion recognition, which combines different modalities for analysis (Lukas Stappen et al., 2021). For effective multimodal emotion recognition, a dataset containing multimodal data labeled with emotions is required (Schmidt et al., 2018).
In this study, we aim to develop a model that predicts stress levels by using emotion recognition results to support counselors in their decision-making. To achieve this, we have created a multimodal dataset specifically designed for analyzing client emotions during online counseling sessions and have conducted evaluations of this dataset. The dataset includes videos of online counseling sessions between laborers and counselors, with objective emotion labels assigned by third parties. Additionally, the dataset contains stress labels derived from questionnaires and counselor assessments. This provides comprehensive data for the development and evaluation of stress prediction models.

Given the current scarcity of multimodal datasets in Japanese that include both emotion and stress labels, this research begins with the creation of such a dataset. This dataset can be applied to develop systems for assessing the mental well-being of workers by analyzing video data, thereby contributing to advancements in managing workers' mental health.

2 Related Works

In this section, we introduce datasets similar to the one constructed in this study.

2.1 MELD

MELD is a multimodal dataset for emotion recognition in conversation. Approximately 13,000 utterances were extracted from 1433 conversations spoken in the TV series “Friends” featuring multiple actors, and each utterance was labeled with an emotion (one of neutral, happiness, surprise, sadness, anger, disgust, or fear). The

labels include audio, image, and text modalities. The labels are assigned by three annotators, and the final label is determined by a majority vote. As a result of allowing re-annotation, a kappa coefficient of 0.43 was achieved. (Poria et al., 2018)

2.2 MuSe: a Multimodal Dataset of Stressed Emotion

MuSe was created to study the multimodal interactions between the presence of stress and emotional expression and the performance of multimodal functions on emotion and stress categorization. It was created to record both college students during and after the test and to make second-by-second predictions about valence and arousal for subjective emotions. (Mimansa et al., 2020)

The differences between the similar data set and this data set are shown in Table 1.

	This dataset	**MELD**	**MuSe**
Contents	**Online counseling**	TV Drama	Single-person speech
Number of speakers	Alone	Multiple people	Alone
Annotation Interval	Consecutive	Speech units	Speech units
Language	**Japanese**	English	English

Table 1: Differences between similar datasets and this dataset.

3 Data Collection

In this study, data were collected through online counseling sessions conducted by counselors using Zoom [1] with Japanese workers. The following section 3.1 describes the video data collection method, and section 3.2 describes the results of the video data collection.

3.1 Video Data Collection Methods

In the online counseling interviews, the counselors conducted semi-structured interviews with a total of 50 clients (workers), each lasting approximately 30 minutes, using Zoom. Before the counseling interview, a questionnaire to evaluate stress was administered. This stress evaluation questionnaire included "quantity of work burden," "quality of work burden," "sleeping hours," "whether they wake up in the middle of the night," "daily working hours," and "life satisfaction" (on a scale of 1 to 10), and participants were asked to answer approximately 150 questions in a choice-type questionnaire. The counseling sessions were conducted in the form of semi-structured interviews, in which the participants were asked a set of questions based on the questionnaire, followed by open-ended questions.

3.2 Processing before showing video to annotators

Only the client's image was included in the video, and the video data was anonymized (i.e., face parts were merged with the average face) so that the annotator assigning the label could not identify the individual client at the time of emotion labeling. A deep learning-based face swapping framework called SimSwap [2] was used for anonymity processing. The audio data of both the counselor and the client were used without anonymization. Figure 1 left shows a part of the counseling video after face exchange using SimSwap.

4 Assigning Annotations

In this study, we annotated the collected data with objective emotion labels to create a multimodal dataset. Section 4.1 describes the annotation method, Section 4.2 describes the annotation results, and Section 4.3 discusses the results, Section 4.4 discusses the correlation between the stress questionnaire and the annotation of emotions.

4.1 Annotation Method

Annotation was performed on the collected data. Seven annotators participated in this experiment, and five of them were randomly selected and assigned to annotate each video. In addition, we instructed the annotators to label emotions without overlooking small changes in emotion, because it was considered that online counseling may not express many emotions when annotating videos.

The annotation method was based on the Russell's circle model (James A. Russell 1980), in which the client's emotional valence (X-coordinate) and arousal level (Y-coordinate) were recorded in one-second increments while watching an anonymously processed online counseling

[1] https://zoom.us/

[2] https://github.com/neuralchen/SimSwap

video, and the coordinates were recorded as objective emotion The labels were obtained as objective emotion labels. The label assignment tool was created using JavaScript and HTML. Figure 1 shows the annotation tool we created.

The online counseling video and Russell's circle model are displayed side by side, and emotion coordinates are captured by mouse operations on the Russell's circle model.

All videos are approximately 30 minutes in length. Annotators can pause/play by clicking the screen during playback. If a mistake is found in labeling, the video can be paused, and the scene can be rewound and corrected using the seek bar below the video.

To prevent the annotator from losing the mouse pointer on the circular map, a different color is used for each quadrant, and the target range for the 16 emotion labels is highlighted. For example, if the client's emotion at a given point in time is determined to be "depressed," move the mouse cursor as shown on the right in Figure 2.

As described above, coordinates were obtained every second by mouse operation on Russell's circle model, and continuous labeling was performed. The results of objective emotion labeling are stored as csv data for each video and each annotator. The X-coordinate (pleasantness-unpleasantness emotional valence), Y-coordinate (arousal level), and the number of seconds (every second) were recorded in this data. We also took care not to label other videos in the middle of a video once the video labeling was started. The system also prevents the user from selecting and watching other videos until the video is finished.

The X-coordinate and Y-coordinate of the emotion labels are assumed to be from -300 to 300 and from -300 to 300, respectively. Figure 3 shows the correspondence between the circular map and the coordinate values.

The points indicated by "●" in the circular map indicate the type and intensity of the client's emotion at that point in time. For example, if the mouse cursor is moved to the coordinate in Figure 3 at 20.0 seconds, the csv data obtained will be [X coordinate, Y coordinate, time] = [250, 50, 20.0].

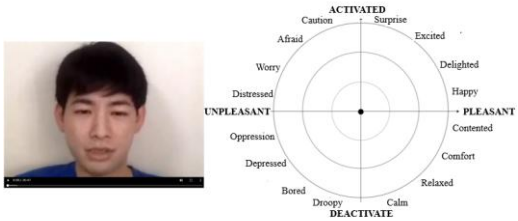
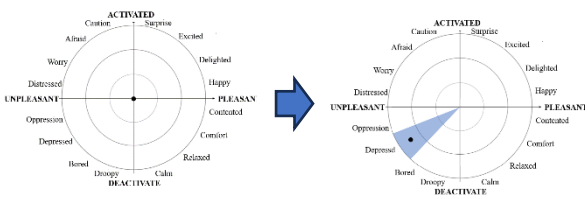


Figure 1: UI for annotation tool to assign emotion labels.



(a) neutral (b) depressed

Figure 2: Annotation Example of "neutral" and "depressed."

4.2 Annotation result and analysis

A total of 410280 labels were assigned by 5 people to all 50 videos. The average number of labels assigned by one person per video was 1641. Figure 4 shows the correspondence between the circle map and the quadrants.

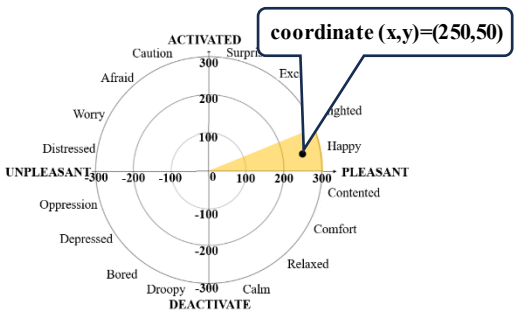


Figure 3: Correspondence between circular map and coordinate values.

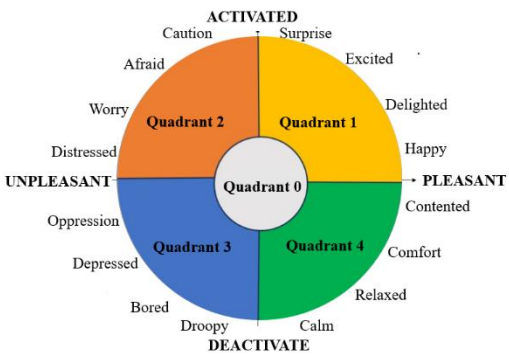


Figure 4: Correspondence between the circle map and the quadrant.

Table 2 shows the number of labels in each quadrant.

Quadrant	Category Labels	Number of label data
Quadrant 0	neutral	248,040
Quadrant 1	pleasure	5,045
Quadrant 2	anger	54,990
Quadrant 3	sadness	69,702
Quadrant 4	enjoyment	32,503

Table 2: Number of label data in each quadrant.

We used Fleiss' kappa coefficient (Fleiss, J. L., 1971) to evaluate the corpus. The Fleiss' kappa coefficient evaluates the reliability of the constructed dataset. In this study, the Fleiss' kappa coefficient, which corresponds to more than one person among the kappa coefficients, was used because the labeling was done by five annotators. The Fleiss' kappa coefficient is a statistic that expresses the degree of agreement, excluding coincidence, for categorical data. In this study, to obtain the values of the kappa coefficients for the objective evaluation of the five annotators, we divided the data into categorical labels with the following 4 levels of granularity. For each division, a threshold of coordinate values was set.

(1) Divide the value of X into 3 parts (threshold: -100, 100)
(2) Divide the value of Y into 3 parts (threshold: -100, 100)
(3) Divide the value of X into 5 parts (threshold: -200, -100,100,200)
(4) Divide the value of Y into 5 parts (threshold: -200, -100,100,200)

Figure 5 shows the categories.

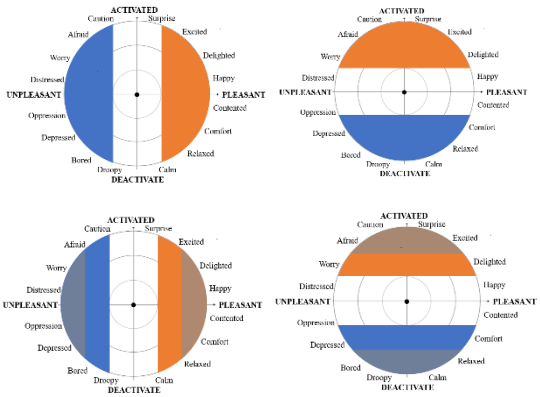


Figure 5: Division into category labels.

The index of Landis et al. The evaluation criteria are shown in Table 3. (Landis, J. R., 1977)

$\kappa<0$	No agreement
$0.00<\kappa<0.20$	Slight
$0.21<\kappa<0.40$	Fair
$0.41<\kappa<0.60$	Moderate
$0.61<\kappa<0.80$	Substantial
$0.81<\kappa<1.00$	Almost perfect

Table 3: Criteria for Fleiss' kappa coefficient.

The Fleiss' kappa coefficient for this dataset was determined. The results are shown in Table 4.

	Kappa coefficient	consistency	concentration
Trisection in x-axis direction	0.149	0.614	0.546
Trisection in y-axis direction	0.016	0.764	0.761
5 divisions in x-axis direction	0.094	0.562	0.515
5 divisions in y-axis direction	0.048	0.442	0.414

Table 4: Average value of Kappa coefficient, agreement, and concentration for 50 videos.

Using the Landis et al. index, the results were “slight” for all categories.

4.3 Discussion of corpus evaluation

Comparing the kappa coefficients for pleasant-unpleasant (x-axis) and activate-deactivate (y-axis), the value for pleasant-unpleasant was higher than that for activate-deactivate. On the other hand, the agreement was higher for activate-deactivate.

This may be because there were few situations in which the level of arousal changed significantly in this counseling session, and most workers moved the mouse pointer up and down less frequently, resulting in higher agreement. Similarly, the concentration level was also higher because the mouse pointer was positioned near the center of the y-axis in more scenes. This is thought to have reduced the value of the Kappa coefficient for activate-deactivate (y-axis direction) in relation to pleasant-unpleasant (x-axis direction). Figure 6 shows a scatter plot of the labels for one video. Colors are assigned to each annotator.

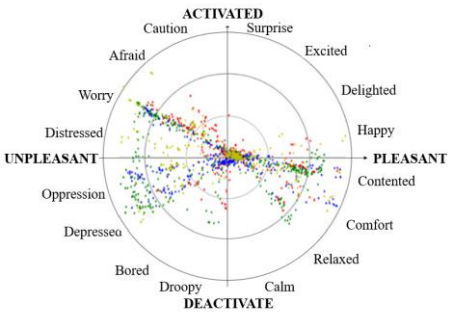


Figure 6: Annotation result.

Calculating kappa coefficients for all of the one-second data tends to result in low values because it does not take into account the aforementioned out-of-sync situations.

4.4 Correlation Analysis Between Emotions and Stress

A correlation analysis was conducted between the mean values of the XY coordinates of Russell's circle model obtained above and the mean values of the probability of occurrence of each emotion calculated from these XY coordinates, and the stress values calculated from the stress questionnaire answered by the participants in advance. The questionnaire consisted of a 57-item occupational stress questionnaire to be answered on a 4-point scale (1~4), with the total score on the BJSQ (highest score 228) representing the participant's self-reported stress value.

Negative correlations were found for the total score of stress values in the questionnaire and the mean score for each item, with the mean value of the X-coordinate of Russell's circle model, the mean value of the probability of occurrence of the feeling of satisfaction, the mean value of the probability of occurrence of the feeling of ease, the mean value of the total probability of occurrence of the feeling in the quadrant 4, and the mean value of the total probability of occurrence of the feeling in the quadrant 1 and the quadrant 4. Negative correlations were found in the mean values.

Conversely, a positive correlation was found for the mean value of the occurrence probability of the emotion depression, the mean value of the sum of the occurrence probabilities of the emotions corresponding to the quadrant 3, and the mean value of the sum of the occurrence probabilities of the emotions corresponding to the quadrants 2 and 3.

From these results, it can be seen that the higher the x-axis (emotional valence), the lower the stress value tends to be. No correlation was observed for the y-axis (arousal level) with stress values. These results suggest that the x-axis (emotional valence) is a particularly important indicator for predicting stress. Figure 7 shows the correlation between stress values and emotions.

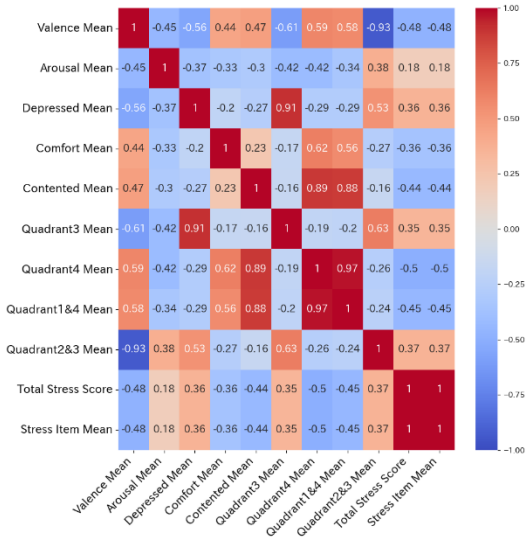


Figure 7: Correlation between emotions and stress levels.

5 Model Construction

Using the constructed dataset, models were built, and emotion estimation was performed. In the following sections, 5.1 describes the feature extraction method, 5.2 describes the feature fusion method, 5.3 describes the model building method, 5.4 describes the emotion estimation results, and 5.5 discusses the results.

5.1 Feature extraction method

The feature extraction procedure is described below as (1) to (4).

(1) Data segmentation method

The data were extracted by excluding scenes in which only the calm label was assigned or in which the subject was not speaking. Specifically, the data were segmented in the silence interval using auditok[3].

[3] https://github.com/amsehili/auditok

(2) Feature extraction from images

Emotion recognition by py-feat (Cheong, J. H., & Xie, S. 2020) is a Python tool for facial expression analysis. It can detect facial expressions (action units, emotions, and facial landmarks) from images and quickly process and analyze them. In this research, emotion recognition results are used as features from images, instead of using features from the entire image. The reason for using the recognition results as features is that facial expression recognition methods are language-independent and are somewhat well established, and because it is possible to eliminate the influence of unnecessary factors such as background. The average value of 30 frames per speech segment was used.

(3) Feature extraction from speech

Acoustic features from wav2vec2.0 (Baevski, A et al., 2020), which has been pre-trained on Japanese speech, are used. wav2vec2.0 learns speech features and builds models using Convolutional Neural Networks (CNN) that have been pre-trained on the voice waveform. At the same time, it is a framework for self-supervised learning for speech representations, achieving high accuracy with only a small transcribed speech data and speech data without correct labels.

(4) Feature extraction from language

For transcribing speech into language, we use a model called NueASR[4], which uses deep learning techniques and is specialized for Japanese speech transcription. It can recognize spoken words with high accuracy. We also use OpenAI's text-embedding-3[5] model. This model is capable of vectorizing linguistic information, supports Japanese, and has the advantage of fast generation speed. The text-embedding-3 model is shown in Figure 8.

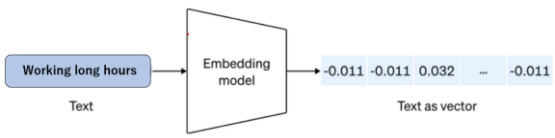


Figure 8: text-embedding-3

5.2 Feature fusion method

In this study, experiments are conducted using a simple concatenation of 1024 dimensions obtained from acoustic features (wav2vec2.0), facial expression recognition results (py-feat), and 1536 dimensions obtained from linguistic features (text-embedding-3). The fusion method used is shown in Figure 9.

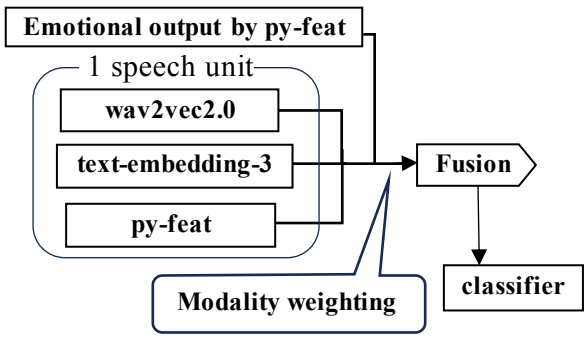


Figure 9: Fusion method used in this study

The main fusion methods of existing research are described below as (1) to (3).

(1) Early Fusion

Early Fusion first combines data from different modalities and then inputs them into a single model. In this method, all modalities are passed to the model at the same time, allowing direct capture of their correlation and interaction. (Jennifer Williams et al., 2018)

(2) Late Fusion

Late Fusion trains separate models for each modality and combines them at the final output stage. This method preserves independence among modalities while allowing complementary information to be leveraged in making the final decision. (Sun, L et al., 2020)

(3) Hybrid Fusion

Hybrid Fusion is a method that combines the advantages of Early Fusion and Late Fusion by fusing some modalities early and others later. This allows for emotion recognition while preserving the important features of each modality. (Cimtay, Y et al., 2020)

[4] https://huggingface.co/rinna/nue-asr

[5] https://huggingface.co/datasets/Qdrant/dbpedia-entities-openai3-text-embedding-3-large-3072-1M

5.3 Model Construction Method

The data was divided into two sets, using 80% of the data for training and 20% for testing. LightGBM (Guolin Ke et al., 2016) was used as the gradient boosting method for training the classifier. This is a type of supervised learning data analysis method that classifies explanatory variables according to an objective variable. The hyperparameters set in this study are as follows. Table 5 shows the hyperparameters used in LightGBM.

Objective	Multiclass
Num_class	16
Num_leaves	62
Learning_rate	0.01
Feature_fraction	0.8
verbose	-1
metric	Multi_logloss
num_boost_round	100

Table 5: Hyperparameters used in LightGBM.

The features of each modal of the speech unit (defined as a segment of speech divided by silent intervals) and the emotional output results of py-feat are input to the LightGBM.

5.4 Emotional Prediction Results

When training the classifier, we assigned a weight to each modality: features from images, features from audio, and features from language. The minimum weight for each feature is 0.2.

Using this weighting, we performed 16 emotion prediction experiments. These 16 emotion assignments are shown in Figure 10.

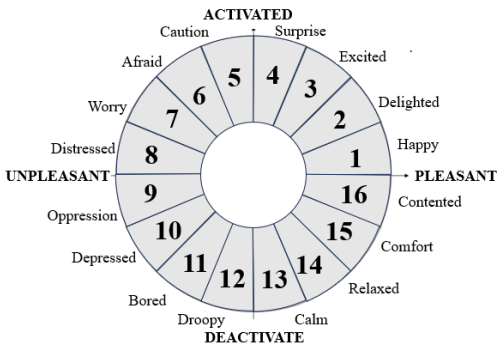


Figure 10: 16 emotions to predict.

Each speech unit is predicted to have one label from the set of 16 emotion labels. The correct label for each speech unit is determined through a majority vote among the five annotators.

Let us denote the py-feat features as "V", the wav2vec2.0 features as "A", and the text-embedding-3 features as "T". Table 6 shows the results of the 16 emotion classifications.

Feature weight	Selection modal	Accuracy
unweighted	V	0.3802
	A	0.4056
	T	0.3068
	V+A	0.4071
	V+T	0.4461
	A+T	0.4416
	V+A+T	0.4266
V=0.2 A=0.2 T=0.6	V+A	0.4236
	V+T	0.4461
	A+T	0.4326
	V+A+T	0.4491
V=0.5 A=0.2 T=0.3	V+A	0.4251
	V+T	0.4446
	A+T	0.4281
	V+A+T	**0.4506**
V=0.3 A=0.4 T=0.3	V+A	0.4251
	V+T	0.4461
	A+T	0.4326
	V+A+T	**0.4521**

Table 6: Results of 16 classification of emotions.

5.5 Discussion of Emotion Prediction Experiments

The results follow previous studies in that accuracy is improved by combining features from images, speech, and language. In the single-modal case, the results using acoustic features showed the best accuracy, followed by facial expression features, and finally linguistic features. Among the overall weightings, the best accuracy was obtained with a weighting of 0.3 for facial features, 0.4 for acoustic features, and 0.3 for linguistic features. The accuracy of 16 emotion recognition was 0.4521.

For this result, the importance of the features was calculated using LightGBM. The top five most important features are shown in Figure 11. (The number of the features is the number of the dimensions entered into the model)

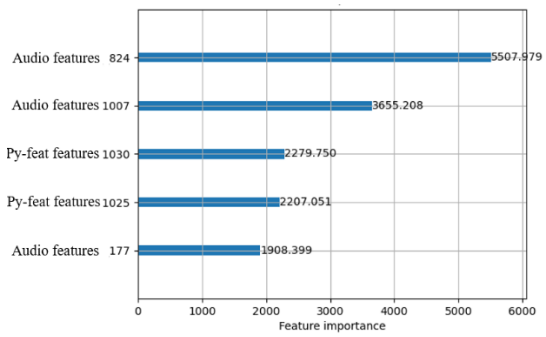


Figure 11: Top 5 most important features.

It was confirmed that the voice feature was more important than the other features. This is consistent with the results of the emotion prediction experiment, in which accuracy was improved when the weight of voice was increased relative to other features in the weighting process.

We hypothesized that speech features are more likely to be expressed to people who have never met before than facial expression or language features.

6 Conclusion

6.1 Summary

In this study, five annotators assigned objective emotion labels to video data (about 30 minutes, 50 people) of stress evaluation interviews conducted with workers using Zoom and constructed a counseling multimodal dataset. Specifically, Russell's circle model was used. An original annotation tool was created, and emotion labels were assigned to the X-coordinate (pleasant - unpleasant) and Y-coordinate (activate - deactivate) every second.

The results of the collected coordinate labels were divided on the pleasant-unpleasant and activate-deactivate axes, and their reliability was evaluated using the Fleiss' kappa coefficient. The results showed that the X-coordinate (pleasant-unpleasant) was higher than the Y-coordinate (activate-deactivate). It is considered that there are differences in response to stimuli (emotional evaluation) and time differences among people. In addition, we used indices such as the Kappa coefficient for each second, but there is room for further investigation as to whether the evaluation for each second is correct or not.

Although we discussed agreement as an objective label, we believe that agreement is difficult to achieve because the task of predicting the client's emotion is subjective in the first place.

In the emotion recognition experiment, we conducted a classification experiment of 16 emotions. Comparing the results of emotion recognition from images, acoustic features, and linguistic features with those from fusion, we found that the accuracy was higher in the fusion case. The accuracy results for emotion recognition in a single modal were 0.3802 for emotion recognition from images, 0.4056 for emotion recognition from acoustic features, and 0.3068 for emotion recognition from linguistic features. The maximum accuracy resulting from the fusion of these features with weights was 0.4521. The weights for each modal were as follows: 0.3 for the emotion recognition results from images, 0.4 for the speech features, and 0.3 for the language features.

6.2 Future Issues

There is a value of stress intensity assigned to each client by counselors and occupational physicians. We would like to compare this value with the annotations and emotion recognition results obtained in this study.

In addition, we would like to analyze the trend of the output of the emotion recognition experiment in a time series and compare it with the results of the annotations and the emotion recognition results obtained in this study.

We would like to see the trend by analyzing the trend of the correct and incorrect parts of the emotion recognition results.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20K12027, JKA and its promotion funds from KEIRIN RACE, and Japan's National Research and Development Agency New Energy and Industrial Technology Development Organization (NEDO)(JPNP20004).

References

Lukas Stappen , Alice Baird , Lukas Christ , Lea Schumann, Benjamin Sertolli, EvaMaria Meßner, Erik Cambria, Guoying Zhao, Bjorn W. Schuller. 2021. *The MuSe 2021 Multimodal Sentiment Analysis Challenge: Sentiment, Emotion,*

PhysiologicalEmotion, and Stress.

Schmidt, P., Reiss, A., Duerichen, R., Marberger, C., and Van Laerhoven, K. 2018. *Introducing WESAD, a multimodal dataset for wearable stress and affect detection. Proceedings of the 20th ACM International Conference on Multimodal Interaction,* pp.400-408.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. 2018. *MELD: A Multimodal Multi-party Dataset for Emotion Recognition in Conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* pp.527-536.

Mimansa Jaiswal, Cristian-Paul Bara, Yuanhang Luo, Mihai Burzo, Rada Mihalcea, and Emily Mower Provost. 2018. *MuSE: a Multimodal Dataset of Stressed Emotion. In Proceedings of the Twelfth Language Resources and Evaluation Conference,* pp.1499-1510.

James A. Russell.1980. A circumplex model of affect. it Journal of Personality and Social Psychology, Vol.39, No.6, pp.1161-1178.

Fleiss, J. L., 1971. Measuring nominal scale agreement among many raters, Psychological Bulletin, 76(5): 378-382.

Shrout, P. E. and Fleiss, J. L. 1979. 1ntraclass correlations: Uses in assessing rater reliability. Psychologica1, Bulletin, 86(2), pp.420-428.

Landis, J. R. and Koch, G. G. 1977. An Application of Hierarchical Kappatype Statistics in the Assessment of Majority Agreement among Multiple Observers.

Cheong, J. H., & Xie, S. 2020. "py-feat: Python Facial Expression Analysis Toolbox." *Journal of Open-Source Software*, 5(47), 2001. DOI: 10.21105/joss.02001.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. 2020. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations." *Advances in Neural Information Processing Systems (NeurIPS)*. Available at arXiv.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu, 2018. *Recognizing emotions in video using multimodal dnn feature fusion , Proceedings of the First Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML),* pp.11-19.

Sun, L.; Lian, Z.; Tao, J.; Liu, B.; Niu, M. 2020. *Multi-modal continuous dimensional emotion recognition using recurrent neural network and self-attention mechanism. In Proceedings of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, Seattle, VA, USA,* pp.27-34.

Cimtay, Y., Ekmekcioglu, E., & Caglar-Ozhan, S. 2020. *Cross-subject multimodal emotion recognition based on hybrid fusion. IEEE Access,* 8, 168865-168878.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang , Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. *LightGBM: A Highly Efficient Gradient Boosting Decision Tree.*