Emotion Estimation Method with Mel-frequency Spectrum, Voice Power Level and Pitch Frequency of Human Voices through CNN Learning Processes

Taiga Haruta[1], Mariko Oda[2], Kohei Arai[3]
Graduate School of Engineering, Kurume Institute of Technology, Kurume City, Japan[1, 2]
Applied AI-Research Laboratory, Kurume Institute of Technology, Kurume City, Japan and Saga University[3]

***Abstract*—Emotion estimation method with Mel-frequency spectrum, voice power level and pitch frequency of human voices through CNN (Convolutional Neural Network) learning processes is proposed. Usually, frequency spectra are used for emotion estimation. The proposed method utilizes not only Mel-frequency spectrum, but also voice pressure level (voice power level) and pitch frequency to improve emotion estimation accuracy. These components are used through CNN learning processes using training samples which are provided by Keio University (emotional speech corpus) together with our own training samples which are collected by our students in emotion estimation processes. In these processes, the target emotion is divided into two categories, confident and non-confident. Through experiments, it is found that the proposed method is superior to the traditional method with only Mel-frequency by 15%.**

Keywords—e-Learning; emotion estimation; Mel-frequency spectrum; fundamental frequency (pitch frequency); sound pressure level (voice power level)

I. Introduction

Since the spread of the new coronavirus, many schools have introduced e-learning, a learning system using information and communication technology (ICT) that allows students to study independently.

Unlike regular classes for many students, e-learning has the advantage that a history is kept for everyone, and learning can proceed according to the student's level of proficiency. However, self-study through e-learning is generally a challenge for students to maintain their motivation to learn. Children and students with intellectual disabilities need psychological support advice beside the disabled child and support to change the questions according to the learner's situation.

Because of “Annual Report on Government Measures for Persons with Disabilities (Summary) 2020”, the total number of mentally retarded persons under the age of 18 is 225,000 [1]. In the Covid-19 pandemic, we believe it is necessary to develop an e-Learning system that considers disabilities to realize "fair, individualized, and optimized learning that leaves no one behind" for a diverse group of children.

As a result of speech change by emotion, it has a significant effect on the continuous speech estimation engine. These effects are caused by the fact that the acoustic models of speech estimation engines currently in general use are designed assuming calm speech. It is necessary to take measures such as switching the acoustic model every time.

Emotion estimation from speech can be divided into methods that use information obtained from the utterance content (context), that is, methods that perform speech estimation and use the results, and methods that use signal processing methods such as prosody, amplitude, and stress. However, it has already been mentioned that the former is difficult to use as the first stage of a large-vocabulary continuous speech estimation system in terms of computational complexity. Therefore, in order to actually estimate emotions, it is realistic to use the latter method of extracting features from emotional speech and estimating emotions. Methods using neural networks and methods using discriminant analysis are conceivable for estimating emotions using the extracted feature values.

It is important that the feature values used in this method reflect the differences between emotions well, and various feature values have been studied so far. It appears in three aspects: prosodic structure such as frequency and formant frequency, amplitude structure such as sentence stress and accent, and temporal structure such as sentence length. After the speech data is transformed into data in the frequency domain by Discrete Fourier Transformation (DFT), it is then subjected to inverse DFT by taking the logarithm. The basic idea of this fundamental frequency extraction method is to select a certain number in descending order and connect these peaks smoothly by Dynamic Programming (DP).

Although the traditional methods for emotion estimation use a combination among the aforementioned features, the most commonly used feature is frequency components (processed from them), and has long been utilized in speech recognition research [2], also the estimation accuracy is not all good. In order to improve the accuracy, the feature combination among frequency components (Mel-Frequency), voice pressure level and pitch frequency is proposed in this paper. One of the targeted applications of the proposed emotion estimation method is a confidence level estimation of students through remote lectures. Actually, interactive lighting using voice emotion recognition is actually being studied [3]. The final goal of this study is the development of learning support system which Artificial Intelligence (AI) plays the role of teacher to understand the learner’s emotions and prompt their self-discipline learning. This paper makes clear how can we determine whether they are confident in their answers from their utterances during learning.

The related research works are described in the following section. After that, the proposed method for emotion estimation with voices is described followed by experiments. Then conclusion is described with some discussions.

II. Related Research Works

Although voice recognition is now world widely available, recognition performance is not good enough for normal conversations. For instance, voice recognition performance of the typical Hidden Markov Model: HMM based method [4] (this is referred to the conventional voice recognition hereafter) with the feature of Formant is less than 50 % when the signal to noise ratio is below 5dB. In other words, voice recognition performance is totally affected by noise. In normal conversation among us, not only voice but also mouth movement is used for recognitions. Mouth movement video analysis makes voice recognition much better performance. The lip-reading method is for improvement of voice recognition performance.

Usually, Hidden Markov Model based method or neural network-based method is used for voice recognitions as well as optical flow [5]-[12] based analysis of the mouth movement videos. Forward direction (from the past to the future) of optical flow is usually used for mouth movement analysis. Voice recognition performance can be improved by adding backward direction (from the future to the past) of optical flow for correction of voice recognition errors through a confirmation of recognized results. In this process, two voice elements are treated as a unit for the proposed backward optical flow. The conventional forward direction of optical flow recognizes by voice element by voice element, though. In order to make sure the recognized results, two voice elements are much easier and efficient manner. This is because transient between voice element and voice element is so important for voice recognitions. This is the basic idea of the proposed lip-reading method.

As for the voice recognition methods, there are the following related research works,

E-learning system which allows students' confidence level evaluation with their voice when they answer to the questions during achievement tests is developed [13]. On the other hand, voice recognition method with mouth movement videos based on forward and backward optical flow is proposed and validated [14]. Mobile device based personalized equalizer for improving hearing capability of human voices for elderly persons is developed and realized [15]. Meanwhile, hearing aid method by equalizing frequency response of phoneme extracted from human voice is proposed and validated [16].

English pronunciation practice system using voice and video recognitions based on optical flow is developed and validated [17]. The eye based domestic helper robot allowing patient to be self-services through voice communications is also developed and validated its usefulness [18].

As mentioned in the previous section, speech features are essential for speech recognition. Recently, in the frequency component, Mel frequency magnitude coefficient was able to recognize emotions with up to 95.25% accuracy on a multi-class support vector machine [19].

Also, Unsupervised feature selection combining Gammatone Cepstral Coefficients (GTCC) and Power Normalized Cepstral Coefficients (PNCC) is used as speech features to recognize emotion from speech in the presence of various noises if the signal-to-noise ratio from -5 dB to 20 dB is 15 dB or higher [20].

After extracting some statistical features from the fundamental frequencies, we applied different AI methods to examine and study whether they have information about the speaker's emotional state, and found that 89.74% for two emotions, 76.14% for a set of three emotions, and 62.99% for a set of four emotions equal accuracy has been achieved [21].

III. Proposed Method

A. Method Configuration

As mentioned above in introduction section, three features combination, frequency components (Mel-Frequency), voice pressure level and pitch frequencywhich are extracted from the studnets voices and from the speech corpus is used in the emotion estimaion. Jupyter notebook provided by anaconda is used. The programming language used is Python. The following Deep Learning used in the emotion estimation, Convolution Neural Network: CNN + Rectified Linear Unit: ReLu + CNN + ReLu + Pooling + CNN + ReLu + CNN + ReLu + Pooling + CNN + ReLu + CNN + ReLu + Pooling + Fully Connected Layers.

B. Voice Data for Deep Learning

There are two datasets of training samples. One is voice data produced by 17 students of Kurume Institute of Technology to verbally answer "yes" or "no" to 20 questions which are shown in Fig. 1(a) in Japanese, including current events, riddles, and simple calculation problems, and used voice data recorded from their voices.

We also used Keio-ESD[1] [22], a speech corpus provided by the Speech Resources Consortium of the National Institute of Informatics, to perform a binary classification of the recorded speech into positive and negative categories.

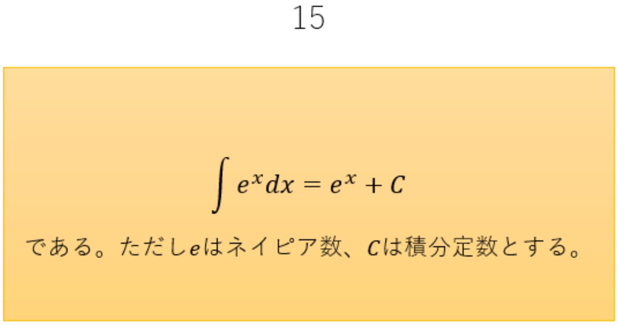
C. Feature Extraction

This time, a comparison was made between a model trained with the Mel-frequency spectrogram alone and a model trained with the fundamental frequency and sound pressure level in addition to it. Therefore, each feature was extracted using Python programming, and the numerical data was stored in csv format and then in a Numpy array.

The resulting Mel-frequency spectrograms, fundamental frequencies, and sound pressure levels were scaled differently for each feature, so the scales were aligned by standardizing the data.

[1] Keio-ESD: Keio University Japanese Emotional Speech Database, http://research.nii.ac.jp/src/index.html, A set of human speech with vocal emotion spoken by a Japanese male speaker and a set of artificial speech that were synthesized by a system that had been developed using the subset of this database for training.

In addition, the number of feature data is proportional to the voice time. This is because the feature data itself is time series data. Therefore, all students recording data was trimmed every 0.5 seconds. In addition, the number of features for each of the 1 voice was matched to the fundamental frequency with the lowest number of features extracted. This is because if the number of features is not aligned, there will be a bias toward the features with the highest number of features during training. Fig. 1(b) is one of the audio waveform data recorded this time. Fig. 1(c) is the Mel-frequency spectrogram extracted from Fig. 1(b). Fig. 1(d) is the fundamental frequency extracted from Fig. 1(b). Fig. 1(e) is the sound pressure level extracted from Fig. 1(b), respectively.

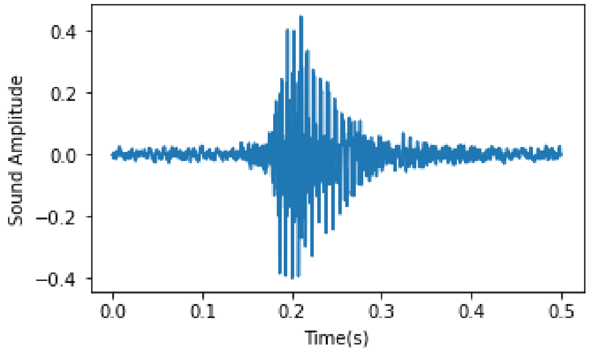


15

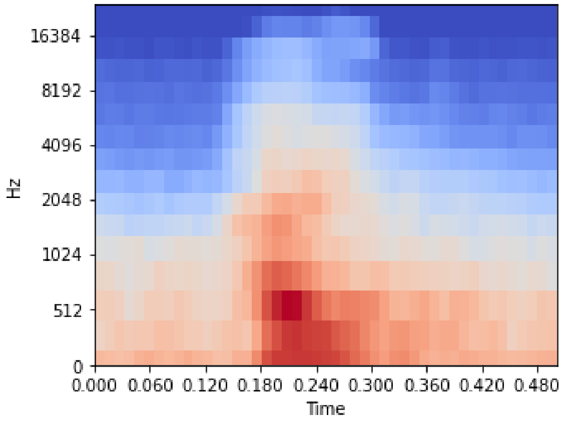
$$\int e^x dx = e^x + C$$

である。ただしeはネイピア数、Cは積分定数とする。

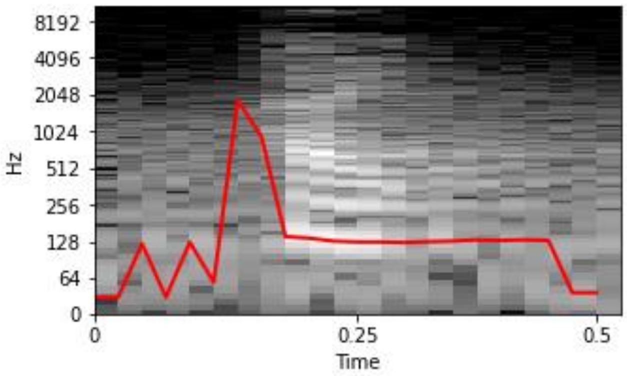
(a) Some of the questions that were asked



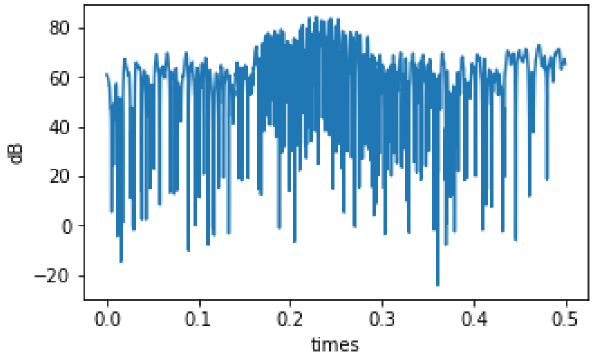
(b) One of the audio waveform data recorded this time.



(c) The Mel-frequency spectrogram extracted from Fig. 1(b)



(d) The fundamental frequency extracted from Fig. 1(b)



(e) The sound pressure level extracted from Fig. 1(b)

Fig. 1. A Portion of the Training Samples Extracted from the Students' Voices.

D. Preparation for Supervised Learning and Learning Model

The speech corpus "Keio-ESD" contains 20 words with 47 different emotions. Among those, we labelled 12 positives: "pleasure", "kind", "gentle", "tolerance", "glad", "admiration", "pride", "love", "satisfaction", "expectation", "happiness", and "like". Then we labelled 12 negatives: "repugnance", "complaint", "disappointment", "sorrowful", "fear", "indifferent", "lamentation", "boredom", "dislike", "Nope." "dejection", and "anxiety". We used the speech for the emotions "expectation," "happiness," "like," "Nope." "Dejection", and "anxiety" as test data and the speech data for the other emotions as training data. However, we did not use the last 20% of the recorded data for training but used it as validation data to measure the goodness of the model parameters after each training.

As for the students' confidence classification of the recorded data, they were asked to answer in advance their confidence after their answers. We then used the recorded data of 13 of the 17 students in our study data as training data. We then used the recordings of the remaining four individuals as test data. We did not use these data for the last 20% of recordings but used them as validation data too.

We used the Function API for CNN with Keras to build our learning model. We trained the Keio-ESD emotion classification 200 times and the student confidence classification 20 times, then evaluated each test data set and compared the results.

IV. Experiemntal Results

A. *Binary Classification of Emotional Speech Corpus into Positive and Negative*

As for the binary classification of speech into positive and negative for the emotional speech corpus, Fig. 2 shows the transition graph of the loss function for each epoch.

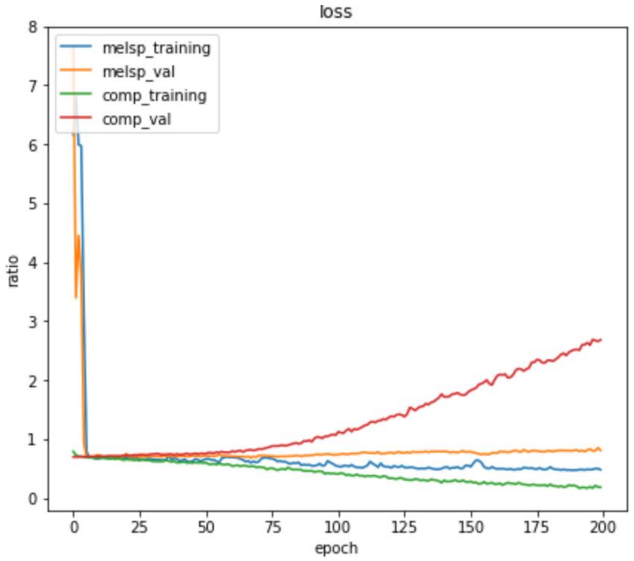


Fig. 2. The Loss Function of Binary Classification of Emotional Speech Corpus into Positive and Negative.

The vertical axis shows the percentage, and the horizontal axis shows the number of epochs. The precedents are discussed below:

- "melsp_training" shows the result of training with only the training data of the Mel-frequency spectrogram in the input layer.
- "melsp_val" shows the results that were checked against the validation data of the Mel-frequency spectrogram after each study.
- "comp_training" shows the result of training with only the training data of the Mel-frequency spectrogram, fundamental frequency, and sound pressure level in the input layer.
- "comp_val" shows the results that were checked against the validation data of the Mel-frequency spectrogram fundamental frequency, and sound pressure level after each study.

Figures are basically rounded to two decimal places.

Regarding the loss function, it was quite large immediately after the Mel-frequency spectrogram was put in the input layer, but after each epoch, melsp_training and melsp_val converged around 0.48 and 0.81, respectively. Concerning comp_training and comp_val, comp_training eventually converged around 0.18, comp_val eventually converged around 2.68.

Fig. 3 is a graph of the percentage of accuracy. With regard to the training data, both melsp_training and comp_training showed an upward trend with each epoch. Finally, melsp_training and comp_val showed 0.76 and 0.93, respectively. In other words, comp_training is about 16% better at learning than melsp_training. As for the validation data, the melsp_val and comp_val finally converged to 0.46 and 0.43, respectively.

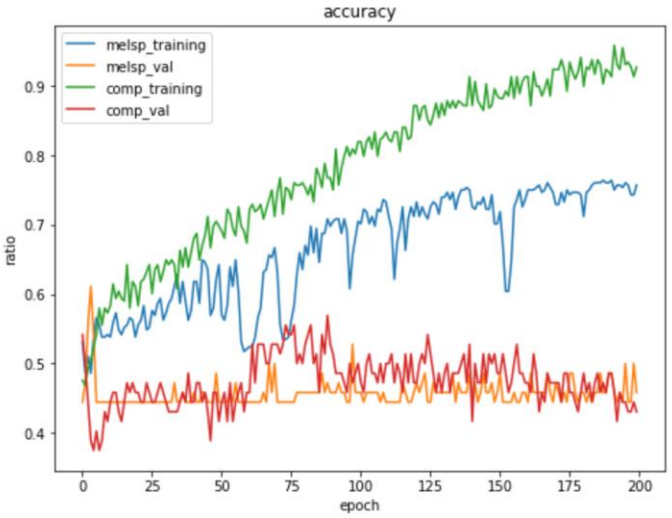


Fig. 3. The Accuracy of Binary Classification of Emotional Speech Corpus into Positive and Negative.

B. *Confidence Classification of the Student Recorded Data*

Fig. 4 is the loss function, and Fig. 5 is the accuracy. Regarding the loss function, as in the previous page, the values of the loss functions for melsp_training and melsp_data were quite large, but after the fourth training, they showed values below 1 and finally converged to 0.69. Regarding the validation data, comp_training eventually converged to 0.55 and comp_val converged to 0.83. comp_val exceeded melsp_val by binary classification of emotional speech corpus into positive and negative same amount.

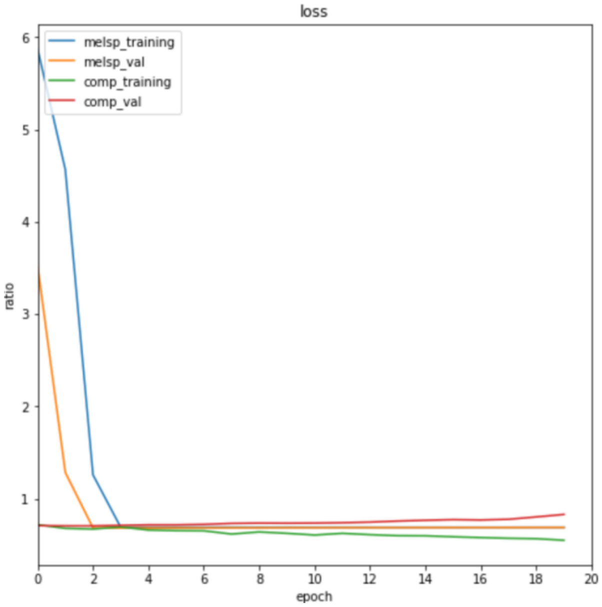


Fig. 4. The Loss Function of Confidence Classification of Student Recorded Data.

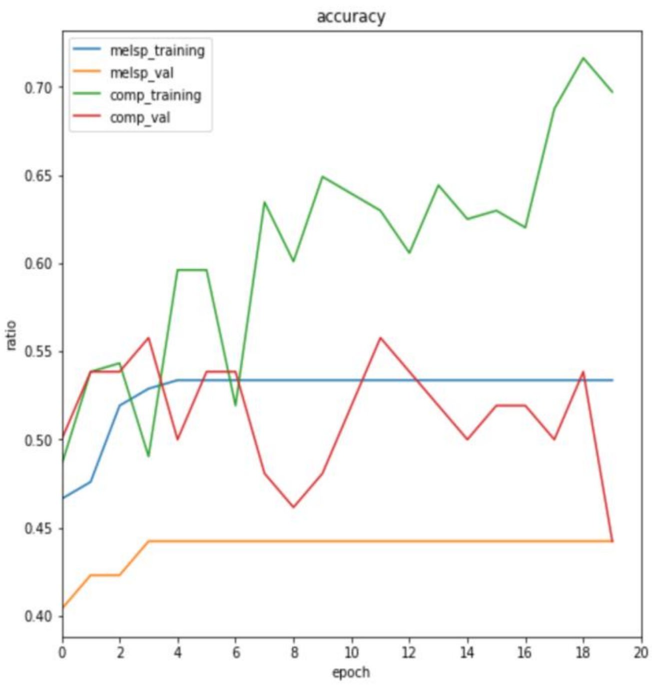


Fig. 5. The Accuracy of Confidence Classification of Student Recorded Data.

With respect to the accuracy, the value of melsp_training remained unchanged from 0.53 after the fifth training, and comp_training increased from 0.49 at the first training to a final value of 0.70. The value of melsp_val also remained unchanged from the fifth study and finally converged to 0.44. Comp_val showed a large difference in the percentage of correct responses from epoch to epoch and finally converged to 0.44.

C. Evaluation with Test Data

We finally evaluated models trained only on the Mel-frequency spectrogram, plus the fundamental frequency and sound pressure level, on test data for each classification. The table of it is Table I and Table II.

TABLE I. EVALUATION WITH TEST DATA (KEIO-ESD)

Binary classification of emotional speech corpus into positive and negative	
Test 1 loss function	1.765992283821106
Test 2 loss function	2.4016263484954834
Test1 accuracy	0.5083333253860474
Test2 accuracy	0.5333333611488342

TABLE II. EVALUATION WITH TEST DATA (CONFIDENCE CLASSIFICATION)

Confidence classification of the student recorded data	
Test 1 loss function	0.6947100162506104
Test 2 loss function	0.7546933889389038
Test1 accuracy	0.48750001192092896
Test2 accuracy	0.637499988079071

Note that Test 1 in the following table shows the results of evaluating the model with only the Mel-frequency spectrogram in the input layer, and Test 2 shows the results of evaluating the model trained with the Mel-frequency spectrogram, fundamental frequency, and sound pressure level.

The loss function for Keio-ESD's emotional voice classification was much higher than 1 for both Test 1 and Test 2, with a difference of about 0.64, while the loss function for the student's confidence classification was also lower than 1, and the difference in the loss function was about 0.06, indicating that the difference between the two functions had narrowed.

On the other hand, Test 2 outperformed Test 1 in both accuracies, with a 15% improvement in accuracy, especially in the classification of students' self-confidence.

V. CONCLUSION

In this study, to develop a learning support system for physically challenged children that applies speech and emotion recognition, which is an applied technology of AI, we attempted to discriminate the confidence of the respondent's speech using deep learning and to classify the emotion of the emotional speech corpus into two emotion values. In this connection, we constructed a model in which only the Mel-frequency spectrogram was learned as the speech feature charge, firstly. Then, not only Mel-frequency spectrum, but also another two types of models were learned: fundamental frequency and sound pressure level. A comparison of the change in the percentage of correct responses between the two models showed that using a model trained with three types of features improved accuracy by up to about 15%. Therefore, it may say that the proposed method is superior to the conventional method with only Mel-frequency spectrum by 15%.

FUTURE RESEARCH WORKS

Further investigation has to be made for improving voice-based emotion recognition accuracy with further experiments by the other testers. The other features related to voice recognition have to be added to the proposed method for improving the recognition accuracy.

ACKNOWLEDGMENT

The authors would like to thank students who participated to our experiments. Also, we used "Keio University Japanese Emotional Speech Database (Keio-ESD)" provided by Speech Resources Consortium, National Institute of Informatics.

REFERENCES

[1] Cabinet Office, Japan "Annual Report on Government Measures for Persons with Disabilities (Summary) 2020 (Japanese) " (2020) P241.

[2] Mengyao Zhang, Kai Cheng, (2021) "Considerations on Speech Feature Extraction for Humming Retrieval", The 83rd National Convention of Information Processing Society of Japan , 251-252,, Information Processing Society of Japan.

[3] Yuichiro Okado, Kumiko Kushiyama, Wataru Kurihara, (2021) "Kotodama : Using Emotion Recognition with Voice Input Interactive lighting production" Proceedings of the Entertainment Computing Symposium (JAPAN), 147-150.

[4] Hongbing Hu, Stephen A. Zahorian, (2010) "Dimensionality Reduction Methods for HMM Phonetic Recognition," ICASSP 2010, Dallas, TX,

http://bingweb.binghamton.edu/~hhu1/paper/Hu2010Dimensionality.pdf .(accessed on September 14 2012).

[5] Huston SJ, Krapp HG (2008). Kurtz, Rafael. ed. "Visuomotor Transformation in the Fly Gaze Stabilization System". PLoS Biology 6 (7): e173. doi:10.1371/journal.pbio.0060173. PMC 2475543. PMID 18651791. http://www.plosbiology.org/article/info:doi/10.1371/journal.pbio.00601 73.(accessed on September 14 2012).

[6] Andrew Burton and John Radford (1978). Thinking in Perspective: Critical Essays in the Study of Thought Processes. Routledge. ISBN 0416-85840-6. http://books.google.com/?id=CSgOAAAAQAAJ&pg=PA77&dq=%22o ptical+flow%22+%22optic+flow%22+date:0-1985.(accessed on September 14 2012).

[7] David H. Warren and Edward R. Strelow (1985). Electronic Spatial Sensing for the Blind: Contributions from Perception. Springer. ISBN 90-247-2689-1. http://books.google.com/?id=I_Hazgqx8QC&pg=PA414&dq=%22optical+flow%22+%22optic+flow %22+date:0-1985.(accessed on September 14 2012).

[8] S. S. Beauchemin , J. L. Barron (1995). The computation of optical flow. ACM New York, USA http://portal.acm.org/ft_gateway.cfm?id=212141&type=pdf&coll=GUI DE&dl=GUIDE&CFID=72158298&CFTOKEN=85078203.(accessed on September 14 2012)

[9] David J. Fleet and Yair Weiss (2006). "Optical Flow Estimation". In Paragios et al.. Handbook of Mathematical Models in Computer Vision. Springer. ISBN 0-387-26371-3. http://www.cs.toronto.edu/~fleet/research/Papers/flowChapter05.pdf.(ac cessed on September 14 2012)

[10] John L. Barron, David J. Fleet, and Steven Beauchemin (1994). "Performance of optical flow techniques". International Journal of Computer Vision (Springer). http://www.cs.toronto.edu/~fleet/research/Papers/ijcv-94.pdf.(accessed on September 14 2012)

[11] B. Glocker, N. Komodakis, G. Tziritas, N. Navab & N. Paragios (2008). Dense Image Registration through MRFs and Efficient Linear Programming. Medical Image Analysis Journal. http://vision.mas.ecp.fr/pub/mian08.pdf.(accessed on September 14 2012)

[12] Christopher M. Brown (1987). Advances in Computer Vision. Lawrence Erlbaum Associates. ISBN 0-89859-648-3. http://books.google.com/?id=c97huisjZYYC&pg=PA133&dq=%22optic+flow%22++%22optical+flow%22. (accessed on September 14 2012)

[13] Kohei Arai, E-learning system which allows students' confidence level evaluation with their voice when they answer to the questions during achievement tests, International Journal of Advanced Computer Science and Applications, 3, 9, 80-84, 2012.

[14] Kohei Arai, Voice recognition method with mouth movement videos based on forward and backward optical flow, International Journal of Advanced Research in Artificial Intelligence, 2, 2, 48-52, 2013.

[15] Kohei Arai, Takuto Konishi, Mobile device based personalized equalizer for improving hearing capability of human voices in particular for elderly persons, International Journal of Advanced Research on Artificial Intelligence, 4, 6, 23-27, 2015.

[16] Kohei Arai, Takuto Konishi, Hearing aid method by equalizing frequency response of phoneme extracted from human voice, International Journal of Advanced Computer Science and Applications IJACSA, 8, 7, 88-93, 2017.

[17] Kohei Arai, Shinji Matsuda and Mariko Oda, English pronounciation practice system using voice and video recognitions based on optical flow, Proceedings of the International Conference on Information Technology Based Higher Education and Training, Kumamoto, 2001.

[18] Kohei Arai, R. Mardiyanto, The eye based domestic helper robot allowing patient to be self-services through voice communications, Proceedings of the 260th conference in Saga of Image and Electronics Engineering Society of Japan, 139-142, 2012.

[19] J. Ancilin , A. Milton, (2021) "Improved speech emotion recognition with Mel frequency magnitude coefficient", Applied Acoustics Volume 179, article 108046

[20] Surekha Reddy Bandela, T. KishoreKumar, (2021) "Unsupervised feature selection and NMF de-noising for robust Speech Emotion Recognition", Applied Acoustics Volume 172, article 107645

[21] Teodora DIMITROVA-GREKOW, Aneta KLIS, Magdalena IGRAS-CYBULSKA (2019), "Speech Emotion Recognition Based on Voice Fundamental Frequency", Archives of Acoustics, 44, 2, pp. 277–286,

[22] Tsuyoshi Moriyama, Shinya Mori, Shinji Ozawa, A Synthesis Method of Emotional Speech Using Subspace Constraints in Prosody, Journal of Information Processing, pp.1181-1191, Vol.50, No.3, 2009.

AUTHORS' PROFILE

Taiga Haruta: He received BE degree in 2022. He also receives the Kurume Institute of Technology President's Award. He is now working on AI-applied e-Learning research and voice recognition in Master's Program at Kurume Institute of Technology.

Mariko Oda: She graduated from the Faculty of Engineering, Saga University in 1992, and completed her master's and doctoral studies at the Graduate School of Engineering, Saga University in 1994 and 2012, respectively. She received Ph.D(Engineering) from Saga University in 2012. She also received the IPSJ Kyushu Section Newcomer Incentive Award.In 1994, she became an assistant professor at the department of engineering in Kurume Institute of Technology; in 2001, a lecturer ; from 2012 to 2014, an associate professor at the same institute; from 2014, an associate professor at Hagoromo university of International studies; from 2017 to 2020, a professor at the Department of Media studies, Hagoromo university of International studies. In 2020, she was appointed Deputy Director and Professor of the Applied of AI Research Institute at Kurume Institute of Technology. She has been in this position up to the present.She is currently working on applied AI research in the fields of education.

Kohei Arai: He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January 1979 to March 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post-Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science in April 1990. He is now an Emeritus Professor of Saga University since 2014. He was a council member for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is a Science Council of Japan Special Member since 2012. He is an Adjunct Professor of University of Arizona, USA since 1998 and is an Adjunct Professor of Nishi-Kyushu University as well as Kurume Institute of Technology/AI Application Laboratory since 2021. He also is Vice Chairman of the Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 77 books and published 678 journal papers as well as 550 conference papers. He received 66 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.html