Journal of Signal Processing, Vol.24, No.6, pp.229-235, November 2020

PAPER

Continuous Audiovisual Emotion Recognition Using Feature Selection and LSTM

Reda Elbarougy[1], Bagus Tris Atmaja[2,3] and Masato Akagi[3]

[1]Faculty of Computer and Information Sciences, Damietta University, New Damietta, Egypt
[2]Department of Engineering Physics, Sepuluh Nopember Institute of Technology, Surabaya 60111, Indonesia
[3]School of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi 923-1292, Japan
E-mail: elbarougy@du.edu.eg, bagus@ep.its.ac.id, akagi@jaist.ac.jp

Abstract **Speech and visual information are the most dominant modalities for a human to perceive emotion. A method of recognizing human emotion from these modalities is proposed by utilizing feature selection and long short-term memory (LSTM) neural networks. A feature selection method based on support vector regression is used to select the relevant features among thousands of features extended from speech and video features via bag-of-X-words. The LSTM neural networks then are trained using a number of selected features and also separately optimized for every emotion dimension. Instead of utterance-level emotion recognition, time-frame-based processing is performed to enable continuous emotion recognition using a database labeled for each time frame. Experimental results reveal that a system with feature selection is more effective for predicting emotion dimensions for a single language than the baseline system without feature selection. The performance is measured in terms of the concordance correlation coefficient obtained by averaging the valence, arousal, and liking dimensions.**

Keywords: continuous emotion recognition, audiovisual, feature selection, LSTM

1. Introduction

Automatic emotion recognition (AER) systems are expected to become an important technology for human–computer interaction as computers become required to be able to communicate with humans naturally [1]. This technology can be embedded in many computer applications as well as in robots to make them sensitive to the user's emotional state [2]. The methodology for AER includes the extraction of emotion-relevant features and classification. Humans detect the emotional state using different modalities such as speech, image, and text. Therefore, a superior emotion recognition system should imitate the human perception of the emotional state by using different types of information from speech, image, and text.

Several features from different modalities should be extracted and combined as an input to the proposed system to accomplish this task. Among many modalities, audio and visual modalities are the most important cues for humans for perceiving emotions. Mehrabian argued that feelings and attitudes from speech communication comprise 38% of vocal/audio information and 55% of visual expression [3]. Combining both audio and visual information will enhance the performance of recognizing and communicating emotion.

Several studies have indeed shown evidence for certain universal attributes for speech [4, 5], music [6, 7], and both [8], not only among individuals of the same culture, but also across cultures. Dang et al., for instance, performed an experiment in which humans had to distinguish between three and six emotions [9]. Their conclusion was that listeners could perceive emotion from speech sound without linguistic information with an accuracy of about 60% in a three-emotion evaluation and an accuracy of about 50% in a six-emotion evaluation.

On the other hand, facial expressions are widely considered as a universal language of emotion [10–12]. The obtained evidence from audio and visual modalities shows that human perception is universal, which means that humans can detect the emotional state regardless of language from multiple modalities. The purpose of this study is to examine whether a mul-

timodal AER system employing different modalities with feature selection and deep neural networks (DNNs) can be used for detecting the emotional state. Emotions also change rapidly as shown by facial and speech features. Hence, continuous short-term emotion detection is more realistic than long-term emotion recognition.

The term of "continuous", as used in this paper's title, besides having a meaning in the temporal domain also has a meaning in emotion theory. Two approaches are commonly used to consider emotions: categorical and dimensional perspectives. Although the first is common in everyday life, the latter is argued to share a more common pattern among human emotions. Since Darwin argued that biological categories, including emotional categories of a species, do not have an essence because of their high variability, Russel and Mehrabian [13] argued that emotional states of people caused by environmental influences could be modeled in the continuous space of pleasure, arousal, and dominance (PAD) model. Instead of dominance, in this paper we use the liking dimension as provided by the dataset [14]. All three emotional attributes are in a continuous space, as well as in continuous time labels. Nevertheless, the term "continuous" in this paper's title represents the temporal domain rather than dimensional spaces.

Although most papers used features from multimodal audiovisual data directly or bag-of-words (BoW) [14–16], we propose to implement feature selection from bag-of-X-words, where X is either audio or video. Feature selection can reduce the computational load of a future system as it reduces the number of features from thousands to a few. By knowing the number of relevant features, it will be easy to achieve real-time emotion recognition by using these few features. To identify the effectiveness of the feature selection method, we compare the systems with and without feature selection along with LSTM networks. Beside optimizing feature selection for each dimension, we evaluate different numbers of selected features to determine which number gives the highest performance.

2. Dataset

A German dataset is used to implement and validate the proposed audiovisual emotion recognition system. This database collects spontaneous and naturalistic interactions consisting of audio and video modalities. All recordings are based on "Sentiment Analysis in the Wild" (SEWA) dataset which consist of remotely recorded human–human interactions. Although the conversation involves two people, only the behaviors of one person are recorded in every recording, as used in Audio/Visual Emotion Challenge (AVEC) 2018. The database contains 130 recordings, which are divided into three different partitions: training, development, and test partitions, as shown in Table 1. This dataset is annotated in three dimensions, namely, arousal, valence, and liking. Gold standard annotations (instead of "true values", since the labels are annotated by humans) were given only for the training and development partitions in continuous degrees and continuous times of arousal, valence, and liking. These labels are given using a hop size of 100 ms, i.e., ten labels per second for each emotion dimension. The training and development data only include recordings from German subjects. However, for the test partition, the German recording does not include any labels for emotion dimensions. The labels for test partition was not provided by organizer of the challenge [14]. The obtained predictions of the test partition were sent to the organizer to obtain the concordance correlation coefficient (CCC) scores reported in this paper. The durations of the recordings in the dataset range from 46 seconds to 3 minutes; the average duration of the recordings is 2.4 minutes.

Table 1 Number of subjects and duration (minutes:seconds) for each partition of recordings

Partitions	Subjects	Labels	Duration
Training	34	✓	93:12
Development	14	✓	37:46
Test	16	-	46:38
Total	64	48	264:36

Although the dataset provides can potentially be used for cross-cultural implementation, we focus on mono-language implementation, i.e., German language implementation, to test our proposed method of feature selection implementation with LSTM networks.

3. Proposed Method

3.1 Baseline features

We use two types of features: audio and visual (video) features. The audio features are mel-frequency cepstral coefficients (MFCCs) and GeMAPS features. For the MFCCs, the first 13 coefficients (0–12), deltas, and deltas-deltas coefficients are extracted; thus, 39 low-level descriptors (LLD) features are used. All audio features are extracted with a window size of 25 ms and a hop size of 10 ms, i.e., 100 audio frames per second. The visual features are 17 facial action units (FAUs) and a confidence feature extracted with OpenFace [17], these features are confidence; facial action units (FAUs) with action intensities of AUs 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 17, 20, 23, 25, 26, and 45, which are extracted with hop size of 20 ms, i.e., 50 frames per second. Both audio and visual features are included

Table 2 Number of audio and video features on baseline feature

Model	LLDs	Bag of words
Audio	23 GeMAPS features	100
	39 MFCCs	100
Visual	1 Confidence and 17 FAUs	100

in the dataset [14] without any modification. Table 2 shows the number of features used in this research.

While the given baseline features provided LLDs extracted based on time frames, it is necessary to aggregate these features to match the given labels extracted per 0.1 seconds. Therefore, the traditional several functional statistics usually extracted from the low-level features on a fixed time (e.g., 0.1 seconds). Instead of using audio and visual features directly, BoW is used as state-of-the-art post-processing for feature extraction. BoW is a document classification technique used commonly in natural language processing.

For text processing, BoW classifies the number of occurrences of a word in a given text document. For audio and visual features, as used in [15, 18], BoW makes a codebook and generates a new feature based on this codebook. First, traditional low-level descriptors such as MFCCs are extracted. Then, using a codebook, LLD vectors are quantized a from single frame across all utterances [18]. The same approach applies for GeMAPS feature set, which extracts 23 features by using openSMILE's `eGeMAPSv01a.conf` file. From both MFCCs and GeMAPS features, we generated 100 new numerical features based on the highest bins of histograms of bag-of-acoustic-word (BoAW) using openXBOW toolkit [18].

For visual features, we extracted 18 facial features and generated the same 100 numerical bag-of-visual-words (BoVW) features. These new features are local image features extracted from images and their general distribution modeled by a histogram [18].

3.2 Feature selection

As described in the previous section, many features were extracted from audio and video, i.e., 100 features from each modality. However, it is not clear which features are the most important for emotion recognition. The baseline system uses all extracted features from each group, i.e., the baseline system measures the impact of the entire group of features on the prediction accuracy, not the impact of each feature individually. Therefore, a feature selection method is required for this purpose. A feature selection method based on a support vector machine, which has been shown to be useful in linear and non-linear classification [19], was proposed to choose n features (expressed as column vector) from 100 features. To measure the impact of each extracted feature individually, each feature was used to predict each emotion dimension using support vector regression (SVR).

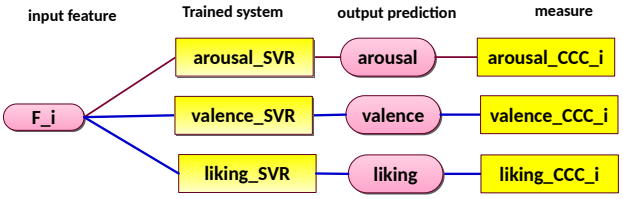

Fig. 1 Use of SVR to select n best features

For the baseline data, features are extracted, as explained in the previous section. Moreover, the gold standard labels for the emotion dimensions of arousal, valence, and liking were evaluated in a listening test using human subjects [14]. To select the set of best features related to emotion dimensions, SVR is used to model the relationship between feature i (F_i) and emotion dimension values (ED). This feature selection was performed by training the SVR model using the training partition and evaluating the trained system using the development partition. To measure the weight of each feature, the concordance correlation coefficient (CCC or ρ_c) between the prediction values of emotion dimensions and the gold-standard values was used. ρ_c is a measure of how well the prediction values of emotion dimensions (Y) compare with a "gold-standard" measurement (X). These gold-standard values is the truth labels provided in the dataset. CCC formulation is given by

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (1)$$

where ρ is the Pearson correlation coefficient (PCC) between the time series prediction and the gold-standard, σ_x^2 and σ_y^2 are the variance of each time series, μ_x and μ_y are the mean values. Therefore, a prediction that is well correlated with the gold standard but shifted in value is penalized proportionally to the deviation. This means that CCC score combines PCC score with the squared difference between the mean of the two compared time series.

Figure 1 shows our SVR method used to select the n best features. From our experimental results, it is known that each feature has a different contribution to different emotion dimensions. We used the average CCC score to obtain n features for all emotion dimensions, where n was manually determined to 6, 10, 15, and 20.

It is difficult to know how many features should be combined to attain the highest prediction accuracy for each emotion dimension. Traditionally, a researcher chooses a threshold value and combines all features that have an impact greater than this specific value. However, the values of CCC are sometimes very small,

which makes it difficult to determine a good threshold. Therefore, we propose a cumulative impact of features algorithm to find the optimal set of acoustic features as described by Algorithm 1. To find the optimal set of features for one emotion dimension, the inputs are the extracted features for the specific groups in decreasing order of the absolute value of CCC. This optimal feature set has 16, 13, and 10 dimensions for arousal, valence, and liking, respectively.

Figure 1 and Algorithm 1 are two integrated components, where in Fig. 1 the impact of each feature is determined, i.e., the contribution of each feature (F_i) for estimating each emotion dimension in terms of the CCC. However, Algorithm 1 is used to calculate the impact and contribution of combining n features using the average CCC score for all emotion dimensions. Finally, Algorithm 1 is used to determine the optimal (feature) set.

Algorithm 1 Cumulative impact of features algorithm to find the optimal set of features for emotion dimension ED_i

Input gold standard values for emotion dimension ED_i for development partition.
Input features sorted by abs(CCC): $f_1, f_2, f_3, ..., f_n$
Input impact CCC_1 of f_1
$Optimal\_Set = \{f_1\}, CCC\_Optimal = CCC_1$
for f_j in $f_1, f_2, f_3, ..., f_n$ **do**
  CCC_j = CCC(Predict ($[Optimal\_Set, f_j]$))
  if $CCC_j > CCC\_Optimal$ **then**
    $CCC\_Optimal = CCC_j$
    $Optimal\_Set = [Optimal\_Set, f_j]$
  end if
end for
return $Optimal\_Set, CCC\_Optimal$

3.3 LSTM networks

We use LSTM networks as the classifier of this emotion recognition system. The LSTM networks used in this research consist of two layers (as suggested in [20]): the first layer contains 128 units, and the second layer contains 64 units. The use of larger units/nodes is expected to improve the performance since the model will learn better than in a small network. We use a batch size of 34 with a 0.001 learning rate. We use unidirectional LSTM since bidirectional LSTM does not contribute significantly in this case. On the other hand, the use of dropout increases the performance. We use a dropout parameter of 0.2. After some experiments, we found that we can limit the maximum number of iterations to 50. These experiments are also used to determine the values of other parameters in LSTM networks.

The use of LSTM networks comes from the idea that humans have the persistence to keep memory long in a short-term period. Humans do not start their thinking from scratch every second. As we read this paper, we understand each word on the basis of our understanding of previous words. We do not throw everything away and start thinking from scratch again. Our thoughts have persistence [21].

Using LSTM as described in [22], we implement an audiovisual AER system by using feature extraction from bag-of-X-words. The selected input features are then trained using German utterances according to their values of arousal, valence, and liking to predict these parameters. We compare the baseline system [14] with 100 BoW features from each audio and visual feature with proposed feature selection method with different numbers of features. We implement the LSTM using the Keras toolkit [23] version 2.3.0 with TensorFlow version 1.14 backend.

For the cost function, we used a custom loss function, namely "CCC loss" instead of the default Keras loss function (MSE). The CCC loss function is defined as

$$CCC\ loss = 1 - CCC \tag{2}$$

The networks are trained to minimize this CCC loss function and maximize the performance in terms of the CCC scores of valence, arousal, and liking. The choice of the loss function is a critical aspect in deep-learning-based pattern recognition since the evaluation of the learning process of the model is based on this metric. Since CCC is used as the evaluation metric, using the CCC loss for the cost function is a straightforward way to achieve a higher CCC score than that using other cost functions.

Figure 2 shows the proposed AER system. As the inputs are audio and visual features that are extracted from audio and video data. BoW extracts 100 features from the previous step within the label-defined time frame, then n features are selected using SVR and fed into LSTM networks. Multitask learning is used to minimize CCC loss from each emotion dimension, as defined in Eg. 2, by summing three losses. The outputs are the predictions of arousal, valence, and liking in continuous times.

4. Results and Discussion

We investigated whether reducing the number of features by a suitable method and processing could improve the performance of a continuous audiovisual emotion recognition system trained using one language. Although the original dataset consists of multiple languages, since the label of emotion dimensions is given only for the German language, the proposed system was trained using only the German language. The baseline system uses all features without investigating the impact of the features used, regardless

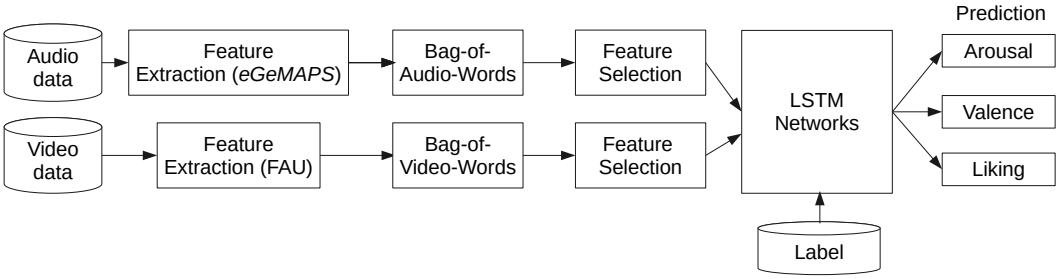

Fig. 2 Proposed continuous audiovisual emotion recognition system

whether they are related to the predicted emotion dimension or not. However, in this paper, the most important acoustic features for the German languages were selected, as explained in the previous section.

To measure the effect of the number of features on the prediction accuracy, the proposed AER system is trained and evaluated using different numbers of features. The results of each emotion dimension using the top 6, 10, 15, and 20 selected features are compared with the optimal set and the baseline features. An optimal set is a number of features tuned for the best CCC score of each emotion dimension, as described by Algorithm 1. The CCC defined in Eq. 1 is used as a metric for evaluating the accuracy of each emotion dimension. However, the performance of the three dimensions was evaluated for each system using CCC. It is still difficult to evaluate the performance of the entire system because the AER system has three outputs. This difficulty is due to the first output (i.e., valence CCC score) is higher than the second output (i.e., arousal CCC score) and third output (i.e., liking CCC score) from the same inputs features. Therefore, it is difficult to determine which system is the best. The average CCC scores for the three emotion dimensions then is used as a measure for the performance of the entire system; this measure is defined by the following equation:

$$CCC_{avg} = \frac{\Sigma\ CCC_i}{n} \tag{3}$$

where n is the number of emotion dimensions, i.e., three (valence, arousal, and liking).

To evaluate the proposed method, the AER system was trained using the training partition and evaluated using the development partition (see Table 1). Table 3 presents the results of evaluation for the development partition using the multi-model of audio (eGeMAPS) and video (FAU) features. From this table, the optimal set gives the best results for arousal and valence. For liking, which is the most difficult emotion dimension to predict, the highest CCC score is obtained by selecting 15 features, meaning that the most features related to liking are from the 15 selected features. Moreover, the average CCC of the optimal set for the three emotion dimensions outperforms other numbers

Table 3 Evaluation results using mono-language case by using the development partition form German language using selected numbers of features from audiovisual modalities (Aro: Arousal, Val: Valence, Lik: Liking, AVG: Average; Baseline consist of 100 features)

Features Set	Aro	Val	Lik	AVG
Baseline [14]	0.552	0.563	0.238	0.451
Selected 6	0.641	0.636	0.278	0.518
Selected 10	0.660	0.620	0.298	0.526
Selected 15	0.622	0.623	**0.314**	0.520
Selected 20	0.616	0.596	0.299	0.504
Optimal Set	**0.678**	**0.654**	0.304	**0.545**

of selected features. The feature selection method in this study was optimized using the development partition of the German language. Evaluation results on the test partition of the German language are shown in Fig. 3. We find that the best performance is achieved using 10 features. The average CCC was increased from 0.382 to 0.416 using this test data. Note that the average CCC score is only used to determine the overall performance among the three emotion attributes as proposed in [14].

Figure 3 shows our best result compared with the baseline. Since the optimal set is difficult to implement in real application, as it has different numbers of features giving the best performance for different emotion dimensions (16, 13, and 10 for arousal, valence, and liking, respectively), we selected 10 features on the basis of CCC score (which gave the second highest performance after the optimal set). Except for arousal, the CCC score from 10 selected features is higher than the baseline score. It can be inferred that the 10 selected features have a high correlation with valence and liking. For arousal, some of the 90 features (100 BoW features − 10 selected features) may have a higher correlation than the 10 selected features, although in the development set the 10 selected features have the highest CCC score after the optimal set.

Finally, we evaluate the model of LSTM networks used as a classifier of the emotion dimensions as a

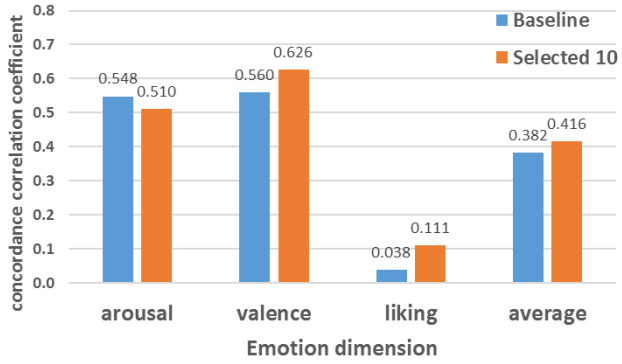

Fig. 3 Comparison between prediction of emotion dimensions for mono language using baseline system and that using proposed system (optimal set) in terms of CCC of test set

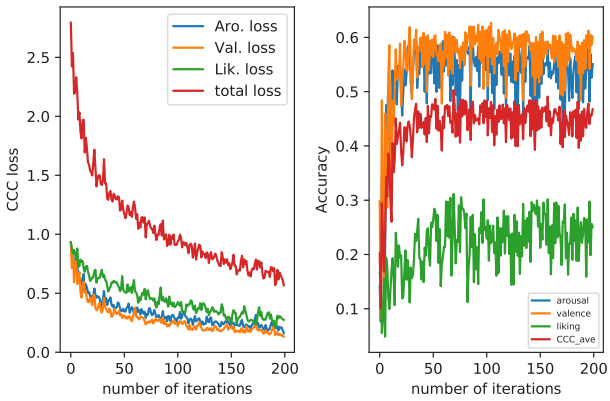

Fig. 4 Training loss and development accuracy (in terms of CCC scores) of LSTM network as a function of the number of iterations (in the case of 10 selected features)

function of the number of iterations. Figure 4 shows the loss and accuracy of the LSTM network, shown in Fig. 2. The loss function used in this model is the CCC loss, as defined in Eq. 2. It is shown that loss function decreases with the increasing number of iterations; it should be close to zero ideally. This can be explained as CCC ideally tend to 1, and the accumulative CCC score for arousal, valence, and liking is 3 (total losses). Hence, the value of each emotion dimension loss $(1 - CCC)$ should tend to zero. In the practical implementation, the CCC loss that is minimized is the total CCC losses from arousal, valence, and liking, which is shown in the left panel of Fig. 4. In the right panel, lines showing the accuracy of valence, arousal, liking, and average CCC are shown. Although the loss functions decreased up to 200 iterations, the accuracies only improved slightly after 50 iterations except for liking. Liking, the dimension that obtained the lowest performance, continued to improve after 50 iterations. This issue (where losses are still going down while the accuracies are stagnant) should be incorporated into future training strategies. The average CCC shows a promising metric as it reflects the performance of all dimensions. This metric should be used as the standard metric in multi-dimensional emotion recognition.

5. Conclusions

A feature selection method was proposed to select relevant features for each emotion dimension from BoW of audio and visual features. The proposed method using SVR was effective for selecting the audiovisual features for each emotion dimension. This technique reduces the number of features from a hundred to a few. Unidirectional LSTM networks using two layers consisting of 64 and 128 units were proposed for estimating the emotion dimensions (arousal, valence, liking). Experimental results show that the proposed method with feature selection can be used to improve the continuous emotion recognition performance from that of a baseline system with a large number of features, as well as reduce its feature dimensions. The evaluation using CCC suggested that the optimum number of features is 10. While this number is chosen manually, future research might investigate the more effective ways to choose the number of most important features automatically. Additionally, since the training loss did not converge in this research, an investigation to evaluate when CCC loss converges in the classifier part is also worthwhile.

References

[1] O. Pierre-Yves: The production and recognition of emotions in speech: Features and algorithms, International Journal of Human-Computer Studies, Vol. 59, Nos. 1–2, pp. 157–183, 2003.

[2] C. M. Lee and S. S. Narayanan: Toward detecting emotions in spoken dialogs, IEEE Trans. on Speech and Audio Processing, Vol. 13, No. 2, pp. 293–303, 2005.

[3] A. Mehrabian et al.: Silent messages, Vol. 8, Wadsworth Belmont, CA, 1971.

[4] R. Banse and K. R. Scherer: Acoustic profiles in vocal emotion expression., Journal of Personality and Social Psychology, Vol. 70, No. 3, p. 614, 1996.

[5] K. R. Scherer, R. Banse, H. G. Wallbott and T. Goldbeck: Vocal cues in emotion encoding and decoding, Motivation and Emotion, Vol. 15, No. 2, pp. 123–148, 1991.

[6] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A. D. Friederici and S. Koelsch: Universal recognition of three basic emotions in music, Current Biology, Vol. 19, No. 7, pp. 573–576, 2009.

[7] P. G. Hunter, E. G. Schellenberg and U. Schimmack: Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions., Psychology of Aesthetics, Creativity, and the Arts, Vol. 4, No. 1, p. 47, 2010.

[8] B. T. Atmaja and M. Akagi: On the differences between song and speech emotion recognition: Effect of feature sets, feature types, and classifiers, arXiv:2004.00200, 2020.

[9] J. Dang, A. Li, D. Erickson, A. Suemitsu, M. Akagi, K. Sakuraba, N. Minematsu and K. Hirose: Comparison of emotion perception among different cultures, Acoustical Science and Technology, Vol. 31, No. 6, pp. 394–402, 2010.

[10] C. Darwin: The expression of the emotions in man and animals, Electronic Text Center, University of Virginia Library, 1872.

[11] P. Ekman: Strong evidence for universals in facial expressions: A reply to Russell's mistaken critique, Psychological Bulletin, Vol. 115, No. 2, pp. 268–287, 1994.

[12] C. E. Izard: Innate and universal facial expressions: evidence from developmental and cross-cultural research., Psychological Bulletin, Vol. 115, No. 2, pp. 288–299, 1994.

[13] A. Mehrabian and J. A. Russell: An approach to environmental psychology., The MIT Press, 1974.

[14] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, E. Çiftçi, H. Güleç, A. A. Salah and M. Pantic: AVEC 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition, Proc. 2018 Audio/Visual Emotion Challenge and Workshop, pp. 3–13, 2018.

[15] Q. Jin, C. Li, S. Chen and H. Wu: Speech emotion recognition with acoustic and lexical features, Proc. 2015 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), pp. 4749–4753, 2015, ISSN 15206149.

[16] F. Eyben, K. Scherer, J. Sundberg, E. And, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan and K. Truong: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective Computing, IEEE Trans. on Affective Computing, Vol. 7, No. 2, pp. 190–202, 2016.

[17] T. Baltrusaitis, A. Zadeh, Y. C. Lim and L.-P. Morency: OpenFace 2.0: Facial behavior analysis toolkit, 13th IEEE Int. Conf. Automatic Face & Gesture Recognition (FG 2018), pp. 59–66, 2018.

[18] M. Schmitt and B. W. Schuller: openXBOW - Introducing the Passau open-source crossmodal bag-of-words toolkit, The Journal of Machine Learning Research, Vol. 18, No. 1, pp. 3370–3374, 2017.

[19] J. Neumann, C. Schnörr and G. Steidl: Combined SVM-based feature selection and classification, Machine Learning, Vol. 61, Nos. 1-3, pp. 129–150, 2005.

[20] A. Karpathy: The unreasonable effectiveness of recurrent neural networks, Andrej Karpathy blog, Vol. 21, p. 23, 2015.

[21] C. Olah: Understanding lstm networks – Colah's blog, http://colah.github.io/posts/2015-08-understanding-lstms/, 2015.

[22] S. Hochreiter and J. J. Urgen Schmidhuber: Long short-term memory, Neural Computation, Vol. 9, No. 8, pp. 1735–1780, 1997.

[23] F. Chollet et al.: Keras, https://keras.io, 2015.

Reda Elsaid Mohamed Elsayed Elbarougy received his B.Sc., and M.Sc., degrees from Mansoura University, Egypt, in May 1997, and February 2006, respectively. Both were in computer science. He was with the Faculty of Science, Mansoura University from 1999 to 2009. In July 2009, he joined the Japan Advanced Institute of Science and Technology (JAIST), Japan, as a Ph.D. student. From September 2014 to August 2019, he was with Mathematics Department, Faculty of Science, Damietta University as an Assistant Professor. In 2017, he was a post-doctoral researcher funded by JSPS to conduct research in JAIST from June 2017 to April 2019. Currently, he is an Assistant Professor in Department of Computer Science, Faculty of Computer and Information Sciences, Damietta University, New Damietta, Egypt, from August 2019 till now. His current research interests include machine learning, artificial intelligence, natural language processing, speech analysis, speech emotion recognition, and synthesis.

Bagus Tris Atmaja received degrees in bachelor and master of engineering physics from the Sepuluh Nopember Institute of Technology in 2009 and 2012, respectively, where he is now employed as a researcher in acoustics. Currently, he is also a Ph.D. student at Japan Advanced Institute of Technology, Nomi, Japan, focusing on speech emotion recognition. His main research interest is speech processing including speech enhancement, source separation, and speech (emotion) recognition.

Masato Akagi received his B.E. from Nagoya Institute of Technology in 1979, and his M.E. and Ph.D. degrees from the Tokyo Institute of Technology in 1981 and 1984. He joined the Electrical Communication Laboratories of Nippon Telegraph and Telephone Corporation (NTT) in 1984. From 1986 to 1990, he worked at the ATR Auditory and Visual Perception Research Laboratories. Since 1992 he has been on the faculty of the School of Information Science of JAIST and is now a full professor. His research interests include speech perception, modeling of speech perception mechanisms in humans, and the signal processing of speech.

(Received November 14, 2019; revised April 30, 2020)