A Comparative Annotator-agreement Analysis of Emotional Speech Corpora

Piyawat Sukhummek[†a)], Jessada Karnjana[††], Sawit Kasuriya[††],
Chai Wutiwiwatchai[††], and Thanaruk Theeramunkong[†]

***Abstract*— This paper proposes three methods for removing or filtering out ambiguous utterances: the filtering based on the first label preference and majority vote, the filtering based on full consensus, and the filtering based on the first label preference and full consensus. We investigate two corpora, which are Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) and Emotional Tagged Corpus on Lakorn (EMOLA). The first corpus is an English language corpus whereas the second one is a Thai language corpus, and both are annotated by six annotators. We primarily study only four emotions, which are anger, happiness, neutral, and sadness. The experimental results show that, once the emotionally ambiguous utterances are removed from a corpus by the proposed methods, and then the corpora are used in training and testing emotion recognition models, the accuracy results improve considerably compared with those of emotion recognition models trained and tested by the original corpora. In the best case, the accuracy improves by 37.47 percents. Also, the proposed methods can considerably improve the reliability of agreement among annotators.**

***Index Terms*—annotator-agreement analysis, inter-annotator reliability measurement, IEMOCAP corpus, EMOLA corpus, HMM-based emotion recognition.**

I. Introduction

EMOTIONS can influence communicative interaction with other people and decision making in many situations. Thus, they are important in human life [1], and understanding in human emotions is one of the essential problems in the cognitive science [2]. There are many areas that study emotions [3], [4], [5]. For example, neuroscientists study them regarding how they can be represented in nervous system activities [6], and psychologists study emotions to explain them and to distinguish feelings [7]. In computer science, research has focused on how we can teach machines to recognize emotions [8] in order to improve the communication between humans and machines.

To develop the emotion recognition technology, we use emotional utterance corpora to construct as well as validate emotion recognition models because these models are normally constructed by using the machine learning approach, and, according to this approach, emotional utterance corpora are required in the training and testing processes [9], [10]. Interestingly, we found from our survey that both processes strongly depend on the corpus used, and hence the performance of the models. Basically, corpora used in this research field contain both emotionally ambiguous and unambiguous utterances. Therefore, the number of emotionally ambiguous utterances in corpora strongly affects evaluation results of the emotion recognition models.

Given a corpus, which is assumed to contain both emotionally ambiguous and unambiguous utterances, this work aims to remove or to filter out emotionally ambiguous ones. In this work, we define the emotionally ambiguous utterance as the utterance to which annotators assign emotion labels differently.

The rest of this paper is organized as follows. Section 2 introduces emotional utterance corpora used in this work and their information. Also, two inter-annotator reliability measures, which are the Scott's Pi and Fleiss' kappa, are briefly reviewed. Our proposed methods are described in Section 3. Then, evaluation and experimental results are given in Section 4. Some remarks and discussion are made in Section 5, and Section 6 summarizes this work.

II. Background

One of the limitations in advancement for the emotion recognition research is the lack of a complete emotional utterance corpus [11]. Many new corpora are designed to overcome such problem and other limitations. Any corpus has four significant key points: scope, naturalness, context, and descriptors [12]. The fist one covers detail information, such as the number of speakers, their spoken languages and genders, the list of emotion categories, and their research goal. The second key point is the speaker types, which are acted, induced, or natural speakers. The third key point is about whether the context has only audio information or has audio and visual information. The last one is about the emotion categories for annotators.

In this work, we use two corpora, which are Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) and Emotional Tagged Corpus on Lakorn (EMOLA), in the evaluation of our proposed methods. The details of these corpora are provided in the following subsection.

A. Corpus Information

The IEMOCAP has 10039 English utterances with ten emotion categories. There are six annotators, and each utterance

† The author is with Sirindhorn International Institute of Technology, Thammasat University, 131 Moo 5, Tiwanon Rd., Bangkadi, Muang, Pathum Thani, 12000, Thailand.

†† The authors are with NECTEC, National Science and Technology Development Agency, 112 Thailand Science Park, Phahonyothin Rd., Klong Luang, Pathum Thani, 12120, Thailand.

a) m5522040376@studentmail.siit.tu.ac.th

Manuscript received April 20, 2016; revised May 01, 2016.

TABLE I: Four key points of the IEMOCAP corpus and the EMOLA corpus.

		IEMOCAP	EMOLA
Scope	Speaker	5 ♂, 5 ♀	21 ♂, 31 ♀
	Language	English	Thai
	Emotion state	Anger, Happiness, Neutral, Sadness, and Frustration	Variety
	Goal	Recognition	Recognition
Naturalness		Acted, Read script	Acted
Context		Audiovisual	Audiovisual
Descriptor		Anger, Happiness, Neutral, Sadness, Excited, Surprise, Frustration, and 3 mores	Anger, Happiness, Neutral, and Sadness

is labeled by three annotators. The default emotion label of each utterance is determined by using the majority vote. The details of the IEMOCAP concerning the four key points are shown in Table I.

The EMOLA has 8987 Thai utterances. There are two emotion levels: primary and secondary emotions. The primary emotion consists of the four following emotions: anger, happiness, neutral, sadness, whereas the other emotions are considered as of the secondary emotion. There are six annotators, and each utterance is labeled by all of them. Note that there is no default emotion label for each utterance. The details of the EMOLA with respect to the four key points are shown in Table I.

B. Inter-annotator Reliability Measurement

In this work, we use two measures, which are the Scott's Pi and Fleiss' kappa, to evaluate the reliability of agreement among annotators. They are briefly introduced in the following subsections.

1) Scott's Pi

The Scott's Pi, which was proposed by William A. Scott [13], can be used to measure the reliability of agreement between two annotators. Let x_{ij} denote the number of utterances that one annotator labels with an emotion category i whereas another annotator labels with an emotion category j. Given that there are N emotion categories and n utterances, the Scott's Pi (π) is calculated by the following equation.

$$\pi = \frac{P - P_e}{1 - P_e}, \tag{1}$$

where

$$P_e = \sum_{r=1}^{N} \left(\frac{1}{2n} \left(\sum_{i=1}^{N} x_{ir} + \sum_{j=1}^{N} x_{rj} \right) \right)^2, \tag{2}$$

and

$$P = \frac{1}{n} \sum_{r=1}^{N} x_{rr}. \tag{3}$$

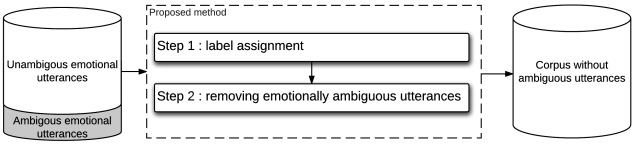

Fig. 1: Framework of the proposed methods.

2) Fleiss' kappa

The Fleiss' kappa is a statistical measure for assessing the reliability of agreement between annotators [14]. Given a corpus with n annotators, N emotional utterances, and k emotion categories including the non-classification one, the Fleiss' kappa κ is defined by the following equation.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{4}$$

where

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{n \cdot (n-1)} \sum_{j=1}^{k} (n_{ij} \cdot (n_{ij} - 1)) \right), \tag{5}$$

$$\bar{P}_e = \sum_{j=1}^{k} \left(\frac{1}{n \cdot N} \sum_{i=1}^{N} n_{ij} \right)^2, \tag{6}$$

and n_{ij} is the number of annotators who label the i^{th} utterance to the j^{th} emotion category. The Fleiss' kappa κ can be interpreted as shown in Table II [15].

TABLE II: Interpretation of the Fleiss' kappa κ.

Fleiss' kappa κ	Agreement Interpretation
less than 0	Poor agreement
0.01 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 1.00	Almost perfect agreement

III. Proposed Methods

Our proposed methods are based on two steps: assigning labels and removing emotionally ambiguous utterances, as depicted in Fig. 1. We propose three different approaches in the label assignment process. Hence, there are three proposed methods, which are the majority vote on the first label, the label consensus, and the first label consensus.

A. Majority Vote on First Label

In the label assignment step, we take only the first emotion label of each utterance and discard the others for each annotator. The reason is that when there are at least two labels assigned to an utterance, it is more likely that the first label is better in representing the emotion based on the utterance. Once each utterance is assigned with one label, we use the majority vote to remove emotionally ambiguous emotion utterances. That is, if there is no majority for an utterance, the utterance is to remove.

TABLE III: Example of utterance labeling with different methods by three annotators. (A:Anger, H:Happiness, N:Neutral, and XXX:Inconclusive)

				Unranked label		First label	
Utterance	**#1**	**#2**	**#3**	**Majority**	**Consensus**	**Majority**	**Consensus**
XXX	A	H	N	✗	✗	✗	✗
A	A	A, N	N	✗	✗	✓	✗
A	A, N	A, N	A, N	✗	✗	✓	✓
N	A, N	H	N	✓	✗	✗	✗
N	A	N	N	✓	✗	✓	✗
N	A, N	H, N	N	✓	✓	✗	✗
N, A	A, N	A, N	N	✓	✓	✓	✗
A	A	A	A	✓	✓	✓	✓

B. Label Consensus

In this method, we use all emotion labels assigned to utterances. To remove emotionally ambiguous emotion utterances, we determine whether there exists an emotion-label consensus or not. If there is no emotion-label consensus in an utterance, the utterance is to remove.

C. First Label Consensus

This method combines two above-mentioned methods. Therefore, it includes removing additional emotion labels and removing emotionally ambiguous utterances. However, an emotionally ambiguous utterance is to remove if there is no emotion-label consensus.

Table III comparatively shows examples of the label assignment steps in each proposed methods.

IV. Simulations and Evaluations

The procedure used in the evaluation of our proposed methods is illustrated in Fig. 2. It consists of four steps as follows. First, the input corpus is modified by our proposed methods. Also, in this step, the Scott's Pi and Fleiss' kappa are used to measure the reliability of agreement among annotators. We compare the measurements of the original and modified corpora. Second, the modified corpus is divided into two groups for training and testing processes. Third, we use the first group to train an emotion recognition model based on the hidden Markov model (HMM). Last, once we obtain the model, it is used in the testing process.

The Scott's Pi values of the original and modified IEMOCAP corpora are compared, as shown in Fig. 3. The Fleiss' kappa values of the original and modified IEMOCAP corpora are compared, as shown in Fig. 5. For the EMOLA corpus, the comparison in Scott's Pi and Fleiss' kappa values between corpus modified with the majority vote and with the consensus are shown in Fig. 4 and Fig. 6, respectively. It should be noted that, since the EMOLA corpus does not have default emotion labels, to compare the Scott's Pi and Fleiss' kappa values we may consider the majority as the default emotion label for each utterance.

The evaluations of the proposed methods regarding the performance of the HMM-based recognition model is measured in accuracy percentage, recall (R), precision (P), F-measure (F) values. The recall is the proportion of the instances correctly against the ground truth judgment of relevance. The precision is the percentage of instances classified as effective that are correct in truth. F-measure is the harmonic mean of the recall and precision [16]. When the proposed methods are applied to the EMOLA corpus, the number of neutrally emotional utterances is reduced greatly. Therefore, in this work, we did not use the modified EMOLA corpus in training the emotion recognition model. The results from the modified IEMOCAP corpus are shown in Table IV and Table V.

TABLE IV: Results from accuracy evaluation of the model trained by the modified IEMOCAP corpus.

	Unranked label		**First label**	
	Majority	**Consensus**	**Majority**	**Consensus**
Testing data	49.72	66.16	49.02	68.35
Training data	52.23	70.61	50.75	72.62

V. Discussion

It can be seen that the accuracy percentages of emotion recognition models trained by the modified corpora improve considerably, compared with those trained by the original ones. The first label consensus method could achieve the best performance compared with the other methods. That is, the accuracy was increased by 37.47 percents. The recall and precision were greatly improved as well. Also, the overall F-measure outperformed the other methods. All proposed methods could improve the reliability of agreement among annotators regarding both measures, significantly.

Although we can achieve the better performance via a reduced corpus using information from emotion labels assigned by annotators, there is still room for improvement. For example, some features of utterances can be combined with information from annotators to remove more emotionally ambiguous utterances from a corpus.

VI. Conclusion

This paper proposed a few efficient methods for analyzing the annotator agreement in order to remove emotionally ambiguous utterances from a corpus. Specifically, given a corpus, which is assumed to contain both emotionally ambiguous and unambiguous utterance, the proposed methods can label them as either ambiguous or unambiguous one. Then, the ambiguous utterances are to remove from the corpus. We proposed three different approaches. The test results show that the accuracy from the corpora modified by proposed methods is better than the original ones. The improvement applies to both corpora. However, when comparing the EMOLA corpus to the IEMOCAP corpus, the improvement is largely noticeable in the IEMOCAP corpus. By removing emotionally utterances from a corpus, this work directly contributes to research in the field of emotion speech recognition.

Acknowledgment

This work was supported under a grant in the TAIST-Tokyo Tech Program.

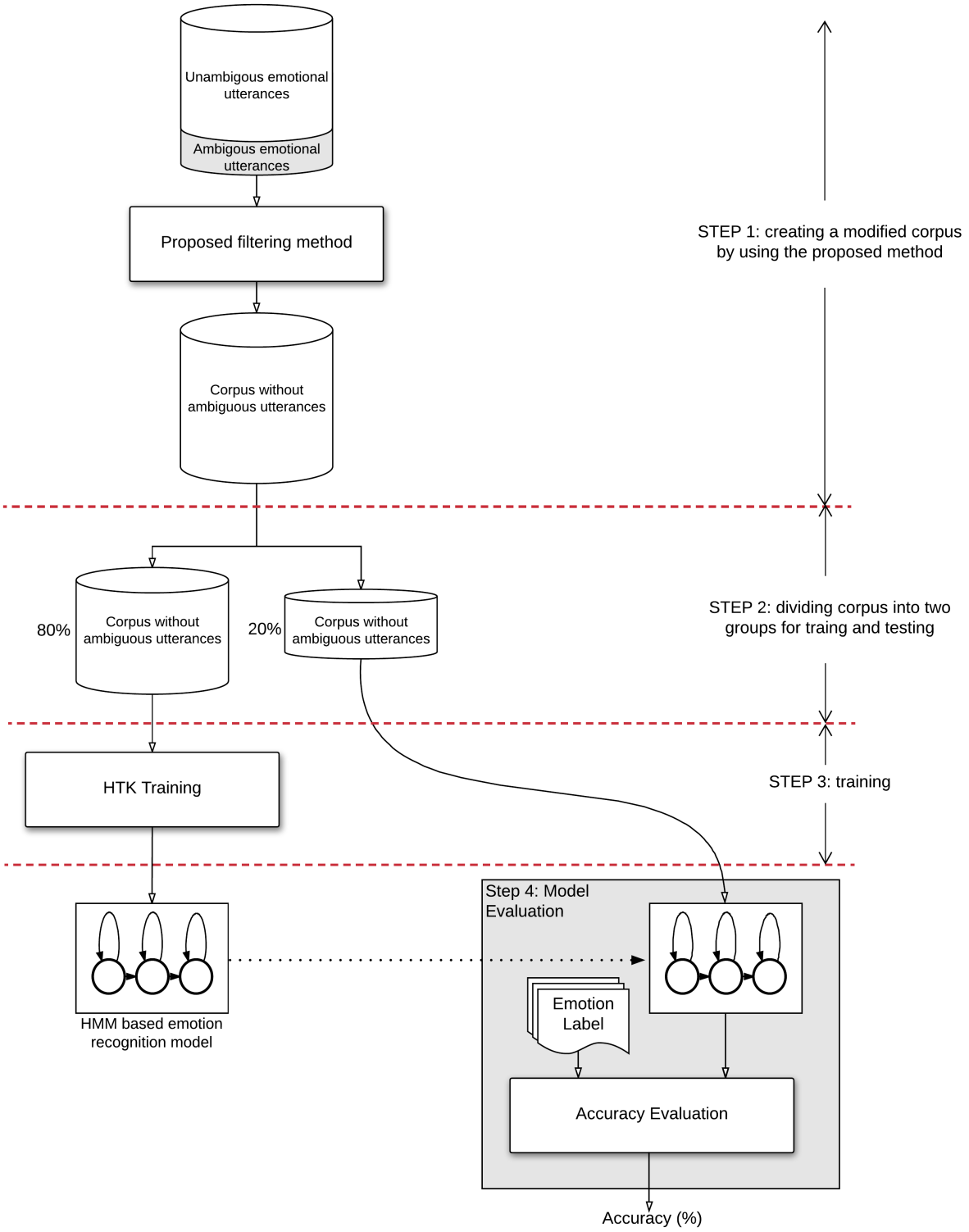

Fig. 2: Experimental procedure.

TABLE V: Performance of the emotion recognition models trained by the modified IEMOCAP corpus on the test dataset measured based on recall (R), precision (P), and F-measure (F).

	Unranked label						**First label**					
	Majority			**Consensus**			**Majority**			**Consensus**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Anger	0.8314	0.5471	0.6599	0.9207	0.7900	0.8504	0.8641	0.5205	0.6496	0.9331	0.8030	0.8632
Happiness	0.2060	0.2604	0.2301	0.1488	0.3830	0.2143	0.1965	0.2535	0.2214	0.1034	0.2344	0.1435
Neutral	0.5557	0.3796	0.4511	0.6051	0.4847	0.5382	0.5902	0.3738	0.4577	0.6859	0.5459	0.6080
Sadness	0.4462	0.7625	0.5630	0.7265	0.7281	0.7273	0.4243	0.7744	0.5482	0.7100	0.7854	0.7458
Average	0.5508	0.4970	0.5000	0.7232	0.6614	0.6855	0.5643	0.4902	0.4948	0.7310	0.6835	0.7013

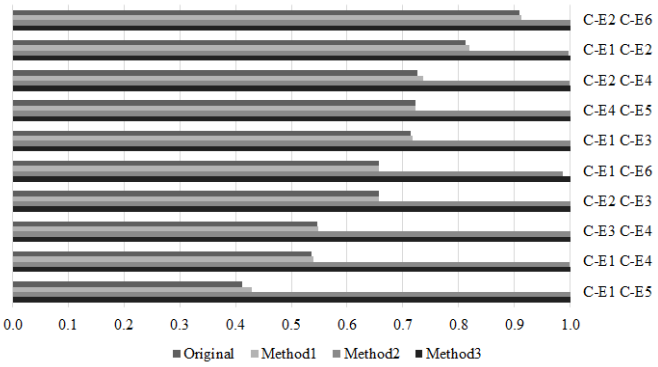

Fig. 3: Scott's Pi value comparison on the IEMOCAP corpus: original corpus, corpus modified with the majority vote on first label (Method1), corpus modified with the label consensus (Method2), and corpus modified with the first label consensus (Method3).

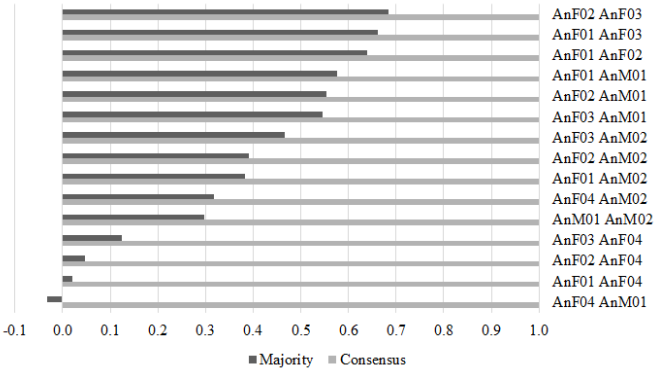

Fig. 4: Scott's Pi value comparison on the EMOLA corpus: corpus modified with the majority vote and corpus modified with consensus.

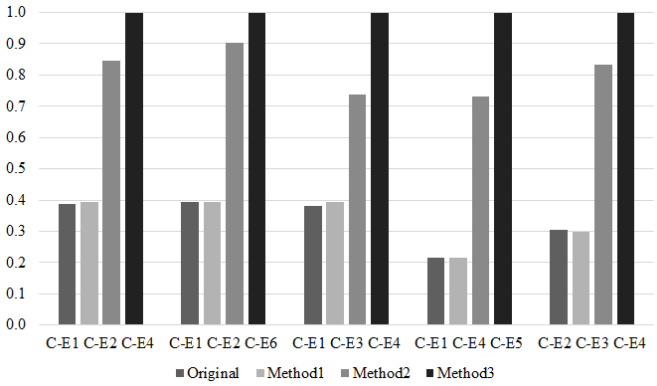

Fig. 5: Fleiss' kappa value comparison on the IEMOCAP corpus: original corpus, corpus modified with the majority vote on first label (Method1), corpus modified with the label consensus (Method2), and corpus modified with the first label consensus (Method3).

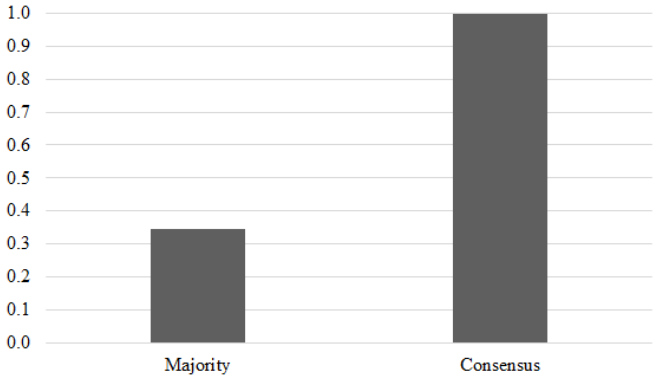

Fig. 6: Fleiss' kappa value comparison on the EMOLA corpus: corpus modified with the majority vote and corpus modified with consensus.

REFERENCES

[1] Hassenzahl, M.: Emotions can be quite ephemeral; we cannot design them. interactions. 11, 5, 46–48 (2004)

[2] Fox, E.: Emotion science cognitive and neuroscientic approaches to understanding human emotions. Palgrave Macmillan, (2008)

[3] Niemic, C.P., Warren, K.: Studies of emotion. A Theoretical and Empirical Review of Psychophysiological Studies of Emotion. 1, 1, 15–19 (2002)

[4] Vul, E., Harris, C., Winkielman, P., Pashler, H.: Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. Perspect Psychol Sci. 4, 3, 274–290 (2009)

[5] Frijda, N.H.: The emotions: Studies in emotion and social interaction. Hist Urbaine. (1986)

[6] Panksepp, J.: Affective neuroscience: The foundations of human and animal emotions. Oxford university press, (2004)

[7] Izard, C.E.: The psychology of emotions. Springer Science & Business Media, (1991)

[8] Pan, Y., Shen, P., Shen, L.: Speech emotion recognition using support vector machine. International Journal of Smart Home 6, 2, 101–108 (2012)

[9] Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Fifteenth Annual Conference of the International Speech Communication Association, (2014)

[10] Lin, Y.L., Wei, G.: Speech emotion recognition based on hmm and svm. In: International Conference on Machine Learning and Cybernetics. vol. 8, pp. 4898–4901. IEEE, (2005)
[11] Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. In: Tenth Annual Conference of the International Speech Communication Association. (2009)
[12] Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a new generation of databases. Speech Commun. 40, 1, 33–60 (2003)
[13] Scott, W.A.: Reliability of content analysis: The case of nominal scale coding. Public opinion quarterly. 321–325 (1955)
[14] Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol Bull. 76, 5, 378 (1971)
[15] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics, 159–174 (1977)
[16] Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. (2011)

Piyawat Sukhummek received his B.Eng. (Computer Engineering) and M.Eng. (Information and Communication Technology for Embedded Systems) from King Mongkut's University of Technology Thonburi and from Sirindhorn International Insitute of Technology, Thammasat University, Thailand, in 2012 and 2017, respectively. His research interests focus on data analysis and speech emotion recognition.

Jessada Karnjana received his B.Eng. (Electronics Engineering), M.Eng. (Microelectronics), Ph.D. (Information Science), and Ph.D. (Engineering and Technology) from King Mongkut's Institute of Technology Ladkrabang, from Asian Institute of Technology, Thailand, from Japan Advanced Institute of Science and Technology, Japan, and from Sirindhorn International Insitute of Technology, Thammasat University, Thailand, in 1999, 2006, 2016, and 2017, respectively. His research interests focus on data analysis and signal processing. He has worked at National Electronics and Computer Technology Center, Thailand, since 1999.

Sawit Kasuriya received his Master degree in Electrical Engineering from Chulalongkorn University. His major research is about signal and natural language processing especially in speech recognition and synthesis. He works for National Electronics and Computer Technology Center (NECTEC) in Thailand since graduated M.Eng.

Chai Wutiwiwatchai received his B.Eng. (1st honor) and M.Eng. degrees of electrical engineering from Thammasat and Chulalongkorn University, Thailand in 1994 and 1997 respectively. He received Ph.D. from Tokyo Institute of Technology in 2004 under the Japanese Governmental scholarship. He is now the Director of Intelligent Informatics Research Unit, National Electronics and Computer Technology Center (NECTEC), Thailand. His research interests include speech processing, natural language processing, and human-machine interaction. His research work includes several international collaborative projects in a wide area of speech and language processing such as multi-lingual speech-to-speech translation under Asian and European countries collaboration and ASEAN languages translation service for ASEAN Economic Community (AEC) 2015. He has positioned as a Vice Chairman of the Asia-Pacific Telecommunity Standardization Program (ASTAP) Expert Group on Speech and Language Processing since 2011. He has been selected one of 66 young leaders shaping Thailand's future by Bangkok Post in 2013. He is now a member of the International Speech Communication Association (ISCA) and the Institute of Electronics, Information and Communication Engineers (IEICE).

Thanaruk Theeramunkong is currently a professor at School of Information, Computer and Communication Technology at Sirindhorn International Institute of Technology (SIIT) at Thammasat University, Bangkok, Thailand. He is also the Program Director of Information and Communication Technology for Embedded Systems (ICTES) at TAIST Tokyo Tech, National Science and Technology Development Agency (NSTDA). He received his bachelor degree in Electric and Electronics Engineering, master and doctoral degrees in Computer Science from Tokyo Institute of Technology. He was a research associate at Japan Advanced Institute of Science and Technology in Japan and a MIS manager at C.P. Seven Eleven Public Co., Ltd. in Thailand. He got several awards, including the Very Good Research Award in engineering field from Thammasat University in 2008, 2009 and 2010. Recently, in 2014, he has received the National Outstanding Researcher Award in the field of Information Technology and Communication Arts. He also got several best paper awards from conferences and societies, including the Japanese Society for Artificial Intelligence, PAKDD workshops, and KICSS. In 2015, he also got a Gold Medal with the Congratulations of the Jury from the 43rd International Exhibition of Inventions of Geneva for the inventions of automatic semantic-based multi-document summarization and application to public hearing. His research interests are natural language processing, data mining, text mining, machine learning and applications to service science. He is also a member of the Steering Committee of the Pacific-Asia Conferences on Knowledge Discovery and Data Mining (PAKDD). He is an associate editor of the Institute of Electronics, Information and Communication Engineers (IEICE). He is the author of more than 40 papers in a number of journals with impact factors and more than 100 conference papers.