TOWARD IMPROVING ESTIMATION ACCURACY OF EMOTION DIMENSIONS IN BILINGUAL SCENARIO BASED ON THREE-LAYERED MODEL

Xingfeng Li and Masato Akagi

Japan Advanced Institute of Science and Technology, Japan

lixingfeng@jaist.ac.jp, akagi@jaist.ac.jp

ABSTRACT

This paper proposes a newly revised three-layered model to improve emotion dimensions (valence, activation) estimation for bilingual scenario, using knowledge of commonalities and differences of human perception among multiple languages. Most of previous systems on speech emotion recognition only worked in each mono-language. However, to construct a generalized emotion recognition system which be able to detect emotions for multiple languages, acoustic features selection and feature normalization among languages remained a topic. In this study, correlated features with emotion dimensions are selected to construct proposed model. To imitate emotion perception across languages, a novel normalization method is addressed by extracting direction and distance from neutral to other emotion in emotion dimensional space. Results show that the proposed system yields mean absolute error reduction rate of 46% and 34% for Japanese and German language respectively over previous system. The proposed system attains estimation performance more comparable to human evaluation on bilingual case.

Index Terms— Emotion dimensions, Fuzzy inference system (FIS), Three-layered model, Emotion recognition in speech.

1. INTRODUCTION

An affective Speech-to-Speech Translation(S2ST) system that has ability to preserve the affective state conveyed in the original speaker's message for converting a emotional spoken utterance in one language to another language plays a critical role for human natural communication across cultures. To produce a colored output with emotional states of the affective S2ST system, automatic speech emotion recognition system is an indispensable component to the affective S2ST system [1] for detecting the emotional sates in the source messages emitted by human beings. A number of previous studies on speech emotion recognition can work well in each mono-language, i.e. changing the source language requires changing the training language for speech emotion recognition system (SERS) [2] [3]. However, several analyses have indeed shown evidence for certain universal attributes for speech, not only among individuals of the same culture, but also across culture. Emotional states can be distinguished from speech even without understanding of that language used [4]. Therefore, developing a generalized speech emotion recognition system that overcomes the retraining issue for multiple languages is full of significance.

To construct the speech emotion recognition system, generally, there are two ways to capture and describe the emotion content in speech: categorical and dimensional approach. Conventional techniques of speech emotion representation focused only on the classification of emotional states as discrete categories such as happy, sad, fear, surprise, and disgust [5]. In everyday interaction, emotional expression always with different intensity depending various situations. However, with a small number of discrete categories, a complex mental state or blended emotions are difficult to handle. Therefore, a dimensional approach in which emotional states are not classified into one of the emotion categories but estimated on a continuous valued scale in a multi-dimensional space for describing human emotion are widely used to reflect the gradient nature of emotion expressions [6].

A number of studies on emotion recognition are reported based on two-layered model using dimensional approach such as by Valster and Schuller 2013, that attempted to estimate the emotion dimensions of valance and arousal from the extracted audio features using Support Vector Machine Regressors with an intersection kernel [7]. However, it is found that the estimation was better for arousal than that for valance. Furthermore, many researchers tried to investigate new acoustic parameters to improve the estimation of valence, as reported by Wu et al. [8] that attempted to estimate the emotion dimensions by combining spectral and prosodic features. However, the obtained results for valence estimation is remained poor.

Human perception, as described by Scherer [9] who adopted a version of Brunswik's lens model originally proposed in 1956 [10] is a multiple layers process. In 2008, Huang and Akagi adopted a three-layered model for human perception. They assumed that human perception for emotional speech is not directly from a change in acoustic layer but rather from a smaller perception which are expressed by adjectives for emotional speech [11]. To precisely esti-

This study was supported by the Grant-in-Aid for Scientific Research (A) (No. 25240026) and the A3 Foresight Program made available by the Japan Society for the Promotion of Science (JSPS).

mate values of emotion dimensions especially for valence, a three-layered model with dimensional approach is adopted by Elbarougy and Akagi in 2012 which consists of three layers: emotion dimensions in the top layer, semantic primitives in the middle layer and acoustic features in the bottom layer. This three-layered model outperforms the conventional system for all emotion dimensions estimation.

In line with these studies, it can be concluded that two-layered perceptual model is poorly to estimate emotion dimensions directly from acoustic features. However, the meaningful and worthwhile two findings are that,(1) dimensional approach is appropriate for representing emotion states,(2) human emotion perception modeled as multiple layers form [9]which emotions are usually described using various adjectives as a bridge not directly from acoustic features. It indicated that three-layered human perception model effectively works for imitating mono-lingual speech emotion perception. However, to construct a common emotion recognition system that be analyzed regardless of the language used, the limitation of retraining issue is still a challenge.

To investigate whether emotional states can be recognized universally or not, Elbarougy proposed a bilingual speech emotion recognition system using dimensional approach [3] [12] derived from a three-layered model adopted by Huang and Akagi [11]. This bilingual perceptual model was constructed with combined information (common acoustic features and common semantic primitives) between two different databases, one in Japanese and the other in German. Moreover, normalization between languages was done in the acoustical feature layer to avoid language independent. However, it is found that this model can only work for several language pairs, applications of normalization and common features selection methods into the bilingual emotional speech recognition system resulted degradation of accuracy in emotional states estimation compared with mono-language cases.

Commonalities and differences of human perception among multiple languages described in [13] on emotion dimensional space (valence-activation space) indicating that direction and distance from neutral to other emotions are similar among languages. Based on this new findings, to overcome the limitations of degradation of accuracy in emotional states estimation on bilingual case after [12], the ultimate goal of this paper is to improve estimation of emotion dimensions for bilingual emotional speech.

To accomplish this, firstly newly relevant features for emotion dimensions are investigated using two different databases simultaneously, secondly, the newly revised three-layered model is constructed by relevant acoustic features in the bottom layer, semantic primitives in the middle layer and emotion dimension in the top layer. Then, Fuzzy Inference System (FIS) is used to connect the elements among three layers as a bridge. Based on the precisely estimated results in emotion dimension layer, the novel normalization method for positions of emotional states in valence and activation space for Japanese and German languages is applied by extracting direction and distance features from neutral to other emotions to improve speech emotion dimensions estimation which can be adopted for multiple languages emotion recognition. Finally, the proposed three-layered system was assessed by comparing the results with previous system after [12].

2. REVIEW OF PREVIOUS RESEARCH

In this section, the previous study on bilingual speech emotion recognition implemented by Elbarougy [12] is reviewed. In that work, two different emotional databases are used. One is Japanese database, the other is German database. The Japanese database is multiple emotions database recorded by Fujitsu Laboratory using single professional actress including five emotional categories: neutral, joy, cold anger, sad and hot anger. There are 20 different sentences, the actress is asked to produce one sentence in neutral voice for once and in each of the other categories for twice. The German database is the Berlin database that covers seven emotional categories, i.e. anger, boredom, disgust, anxiety, happiness, sadness, and neutral [14]. Ten professional German actors (five females and five males) spoken ten sentences with an emotionally neutral content in seven different emotions. Eventually, 340 utterances are collected from two emotional databases in which an equal distribution of four similar categories are collected as follow, 50 neutral, 50 sad, 50 angry, 50 happy, totally 200 utterances from the Berlin database and 20 neutral, 40 joy, 40 hot anger, 40 sad, a total of 140 utterances from Japanese database. The model constructed by Elbarougy using a three-layered model with emotion dimensions in the top layer, semantic primitives in the middle layer and acoustic features in the bottom layer. All used 21 extracted acoustic features, 17 evaluated semantic primitives and 2 evaluated emotion dimensions are all following up from [12].

The previous system proposed by Elbarougy indicates that emotional states can be recognized universally. This model can work for language pair, i.e. one in Japanese and the other in German. However, motivated of building a universal speech emotion recognition system, this model addressed two aspects of limitations, the first one is the selection method of common acoustic features, the second difficulty is the normalization method between two different languages in the acoustic features layer.

In the case of acoustic features selection, to construct a bilingual speech emotion perceptual model that works for Japanese and German languages, firstly, Elbarougy constructed a perceptual three-layered model individually for each dimensions for the two databases. Then the common acoustic features between two languages were selected to constitute the bottom layer. Moreover, the common semantic primitives between the two-languages were selected as semantic primitives for the bilingual case. While constructing a common speech emotion recognition system across languages, it is impractical to construct a three-layered perceptual model individually for each language to select common features in multiple languages cases. This is because there is no guarantee of numerous highly correlated features for each dimension.

In the case of normalization method between languages, the proposed bilingual speech emotion recognition system originated from Elbarougy's study was validated by training the system using one language, and testing using the second language. For instance, to estimate emotion dimensions for Japanese from German, the acoustic features, semantic primitives and emotion dimensions for German database were used to train this system, then the trained system is used to estimate emotion dimensions for Japanese database, to avoid language independent. The tested acoustic features from Japanese language should be normalized by dividing the values of Japanese acoustic features by the mean value of neutral utterances in German database for all Japanese acoustic features. Lastly, the Japanese normalized acoustic features are used as input to the trained system to estimate emotion dimensions for Japanese, and vice-versa. However, actually the normalization method in the acoustic features layer for avoiding language dependency are difficult to accurately predict positions in the dimensional space because elements in three layers are connected nonlinearly by FIS.

Due to these limitations, the emotion dimensions are difficult to estimate based on the model using normalized common acoustic features proposed by Elbarougy.

3. BILINGUAL SPEECH EMOTION RECOGNITION SYSTEM

3.1. Acoustic Features Selection

Newly correlated acoustic features to emotion dimensions of the proposed bilingual emotion perceptual model are collected in this section. For each emotion dimension, selected acoustic features are considered to be the features most relevant to the used dimension in the top layer. To accomplish this task, next two procedures are carried out.

Firstly, correlations between the elements of the top layer (emotion dimension) and the middle layer (semantic primitive) were calculated using two databases simultaneously by Eq.(1) as follow: let $x^{(i)} = \{x_n^{(i)}\}(n = 1, 2, \dots, N)$ be the sequence of the values of the i^{th} emotion dimension rated with the listening test, $i \in \{Valence, Activation\}$, where N is the number of utterances in our database ($N = 340$ totally for Japanese and German databases). Moreover, let $s^{(j)} = \{s_n^{(j)}\}(n = 1, 2, \dots, N)$ be the sequence of the values of the j^{th} semantic primitive rated with listening tests, $j \in \{Bright, Dark, \dots, Slow\}$, where N is the number of utterances in our database. Then, the correlation coefficient $R_j^{(i)}$ between the semantic primitive $s^{(j)}$ and the emotion dimension $x^{(i)}$ can be determined by the following equation:

$$R_j^{(i)} = \frac{\sum_{n=1}^{N}(s_{j,n} - \overline{s_j})(x_n^{(i)} - \overline{x}^{(i)})}{\sqrt{\sum_{n=1}^{N}(s_{j,n} - \overline{s_j})^2}\sqrt{\sum_{n=1}^{N}(x_n^{(i)} - \overline{x}^{(i)})^2}} \qquad (1)$$

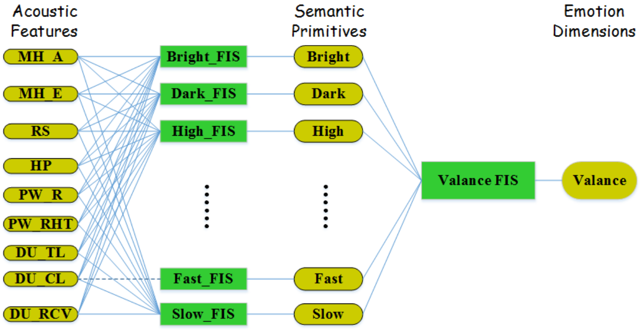


Fig. 1. Bilingual language: valence perceptual model

where $\overline{s_j}$, and $\overline{x}^{(i)}$ are the arithmetic means of the semantic primitive and emotion dimension respectively. Correlation coefficients between semantic primitives and emotion dimensions for the bilingual emotion perceptual model are shown in Table 1. The semantic primitives with correlation coefficients greater than 0.45 are chosen as highly correlated semantic primitives for each emotion dimension to describe this dimension.

Subsequently, the correlation coefficients between elements of the middle layer (semantic primitive) and the bottom layer (acoustic feature) are calculated using equation similar to Eq.(1). In a similar way, the acoustic features with correlation coefficients greater than 0.45 are selected as highly correlated acoustic features. Table 2 lists only 9 acoustic features, which yield a significant correlation with semantic primitives. Based on this acoustic features selection method, finally, a perceptual three-layered model was constructed for each emotion dimension. For example, Fig.1 illustrates the valence perceptual model of the proposed bilingual case.

To implement the three-layered model, an Adaptive-Network-based Fuzzy Inference System is used to connect the elements of the proposed system [15]. To estimate emotion dimensions, a bottom-up estimation method is used. For instance, to estimate valence dimension, 17 semantic primitives in the middle layer are evaluated from 9 acoustic features using 17 FISs, followed by estimating the degree of valence dimension from 17 estimated adjectives by another FIS. In a similar way, activation dimension is evaluated using FIS for each semantic primitive and one FIS for activation dimension.

3.2. Bilingual Language normalization

Commonalities and differences of human perception among multiple languages have been investigated on dimensional space by carrying out listening tests using subjects from different countries in [13]. As shown in Fig.2, it indicated that directions and distances from neutral to other emotions in the emotion dimensional space are similar among languages. Based on the highly precision estimated values in dimensional space from the newly revised three-layered model, in this study, a novel normalization method is presented to realistically imitate human emotion perception by adopting the direction and distance features to recognize emotional states for

Table 1. Bilingual Language: Correlation coefficients between semantic primitives and emotion dimensions. Definitions of the semantic primitives are after [11].

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	Valence	**0.7**	**-0.6**	**0.5**	**-0.5**	-0.1	-0.2	-0.1	0.0	0.1	-0.1	**-0.8**	**0.8**	-0.1	-0.2	0.0	0.2	-0.3	6
2	Activation	**0.7**	**-0.9**	**0.8**	**-0.8**	**0.9**	**-0.9**	**-0.9**	**0.9**	**0.9**	**-0.7**	**-0.5**	**0.5**	**0.9**	**-0.9**	**0.9**	**0.8**	**-0.6**	17
	#	2	2	2	2	1	1	1	1	1	1	2	2	1	1	1	1	1	24

Table 2. Bilingual Language: Correlation coefficients between acoustic features and semantic primitives. Definitions of the acoustic features are after [3].

m		Bright	Dark	High	Low	Strong	Weak	Calm	Unstable	Well-modulated	Monotonous	Heavy	Clear	Noisy	Quiet	Sharp	Fast	Slow	#
1	MH_A	**-0.7**	**0.8**	**-0.7**	**0.7**	**-0.8**	**0.8**	**0.7**	**-0.8**	**-0.7**	**0.6**	**0.5**	-0.4	**-0.8**	**0.7**	**-0.8**	**-0.6**	**0.5**	16
2	MH_E	-0.3	0.4	-0.3	0.4	-0.4	**0.5**	**0.5**	-0.4	-0.3	0.3	0.3	-0.2	-0.4	**0.5**	-0.4	-0.3	0.3	3
3	F0_RS	**0.6**	**-0.8**	**0.7**	**-0.9**	**0.5**	**-0.7**	**-0.8**	**0.6**	**0.6**	**-0.7**	**-0.7**	0.4	**0.6**	**-0.8**	**0.5**	0.3	**-0.5**	15
4	F0_HP	**0.5**	**-0.7**	**0.7**	**-0.8**	**0.5**	**-0.6**	**-0.8**	**0.6**	**0.6**	**-0.7**	**-0.7**	0.3	**0.6**	**-0.7**	**0.5**	0.3	**-0.5**	15
5	PW_R	0.4	**-0.6**	**0.5**	**-0.6**	0.4	**-0.6**	**-0.6**	**0.5**	**0.5**	**-0.6**	**-0.5**	0.3	**0.5**	**-0.7**	0.4	0.2	-0.4	11
6	PW_RHT	0.0	-0.2	0.2	-0.2	**0.5**	-0.4	**-0.5**	**0.5**	**0.5**	-0.4	0.1	-0.2	**0.5**	-0.4	**0.5**	0.3	-0.2	6
7	DU_TL	-0.3	0.3	-0.3	0.2	-0.3	0.3	0.1	-0.2	-0.2	0.0	0.1	-0.3	-0.3	0.2	-0.3	**-0.5**	0.3	1
8	DU_CL	-0.4	0.4	-0.4	0.4	-0.4	0.4	0.3	-0.3	-0.3	0.2	0.3	-0.3	-0.4	0.4	-0.4	**-0.5**	0.4	1
9	DU_RCV	-0.4	**0.6**	-0.4	**0.5**	-0.3	**0.5**	**0.5**	-0.3	-0.3	0.4	**0.5**	-0.3	-0.4	**0.6**	-0.3	-0.2	0.4	6
	#	3	5	4	5	4	6	7	5	5	4	5	0	5	6	4	3	3	74

multiple languages. Firstly, the angle between the direction from neutral to emotional state and the horizontal directions towards positive valence are extracted using following Eq. (2)

$$angle = arctan(\frac{y_E - y_N}{x_E - x_N}) \tag{2}$$

where (x_E, y_E) is the center position of the emotional state E, and (x_N, y_N) is that of the neutral state N. Secondly, the degree from neutral to emotional state in dimensional space is computed by Euclidean-distance in Eq. (3).

$$d(E, N) = \sqrt{(x_E - x_N)^2 + (y_E - y_N)^2} \tag{3}$$

The definition of the parameters are the same as in Eq. (2). Using these two extracted features, the recognition accuracy of this proposed bilingual speech emotion recognition system are evaluated in the next section.

4. SYSTEM EVALUATION

In this section, the evaluation results of the proposed system are presented. To investigate how effectively the proposed system improves emotion dimensions estimation, performance of the proposed system was compared with that of the previous system in [3] [12] on two aspects. The first one is results evaluation for emotion dimension estimation in which Mean Absolute Error (MAE) is used as a metric. The second is results evaluation for emotion classification in which emotion recognition accuracy is used as a metric.

The distribution of emotion dimension estimation for all utterances for both Japanese and German databases are compared directly in the valence-activation space as shown in Fig. 3. Figure3 (a) shows the responses in listening tests by human subjects for all utterances in Japanese and German databases.

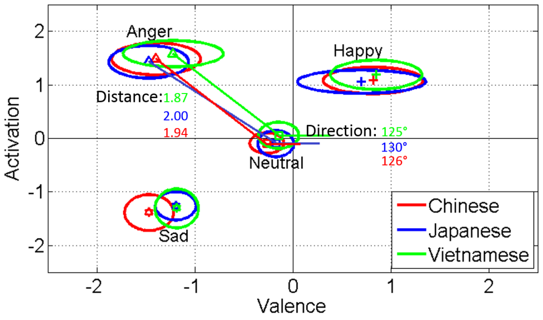
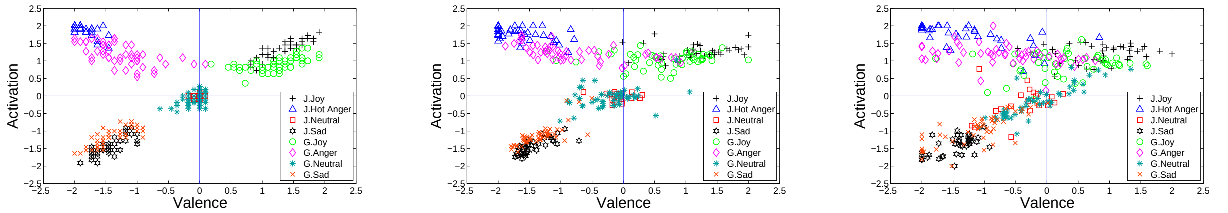


Fig. 2. Positions of emotional states on Valence-Activation emotion space from [13]. Lines from Neutral to Anger indicate directions and distances for three subject groups.



(a) Human responses.

(b) Estimated by the proposed system.

(c) Estimated by Elbarougy's system [12].

Fig. 3. The distribution of the speech utterances in the Valence-Activation space.

Figures 3(b) and (c) show the estimated results by imitating human perception using the proposed system and Elbarougy's system respectively. The results of emotion dimensions from Elbarougy's system in Fig. 3(c) is not able to be accurately estimated, while using the proposed system the estimation accuracy of emotion dimension is improved and comparable to human responses. Based on these estimated results, the proposed system is assessed as follows.

In consideration of the different number of features used in the proposed system and Elbarougy's system, for a rational and logical comparison between these models. Two more systems were trained using different groups of acoustic features and semantic primitives: one system is constructed following Elbarougy's method [12], in which the system was trained by one language and tested by another, using 9 acoustic features (AF) and 17 semantic primitives the same as features used in this paper, the most important key is that the used 9 acoustic features in the bottom layer should be normalized in Elbarougy's way. The other system is constructed using the proposed normalization method in this paper by common features just the same as which selected in Elbarougy's system, this system is trained using two languages simultaneously, moreover, normalization between languages is performed in the emotional space. The above two system were named Elbarougy_9AF and Proposed_commonAF respectively.

4.1. Results Evaluation from Emotion Dimension Estimation

In order to assess the performance of the system for emotion dimension estimation, MAE between the predicted values of emotion dimensions and the corresponding average values responded by human subjects is calculated according to Eq. (4).

$$MAE^{(i)} = \frac{\sum_{i=1}^{n}\left|\widehat{x}_n^{(i)} - x_n^{(i)}\right|}{N} \qquad (4)$$

where $i \in \{Valence, Activation\}$, $\widehat{x}_n^{(i)}$ is output of the proposed system, and $x_n^{(i)}$ is the human responses by using human subjects. The MAEs of Japanese and German databases using the four systems are presented in Fig. 4. In the case of Japanese, the MAE of the proposed system compared with that of Elbarougy's system is reduced from 0.26 to 0.13 for valence and from 0.14 to 0.09 for activation. For German language, the MAE decreased from 0.33 to 0.23, and from 0.16 to 0.10 for valence and activation respectively using the proposed system. The results indicate that the proposed system can precisely estimate emotion dimensions for bilingual languages compared with the previous work by using normalized acoustic features as input parameters.

4.2. Results Evaluation from Emotion Classification

The proposed system is evaluated on speech emotion recognition accuracy. Each point in emotion dimensional space can be mapped into an emotional category. In this part, All 340 utterances were divided into 10 groups, 10-fold cross-validation is applied for training and testing the SVM classifier to map the estimated results into emotion categories. Based on the new normalization method on emotion dimensions proposed in this paper, extracted direction and degree features described in Eq. (2) and(3) are used as training and testing data of the SVM system to map extracted results into categories for evaluating the proposed system. The results are shown in Table 3 for Japanese language case, and in Table 4 for German language case.

The emotion classification accuracies listed in the above two tables correspond to the MAEs for different systems. As represented in Table 3, the average recognition rates for Japanese using the proposed system case is 96% which increased from 89% by using the previous system. The results in Table 4 for German language recognition indicates that the average recognition rates increased from 78% in the case of previous model to 85% in the case of the system trained using proposed method.

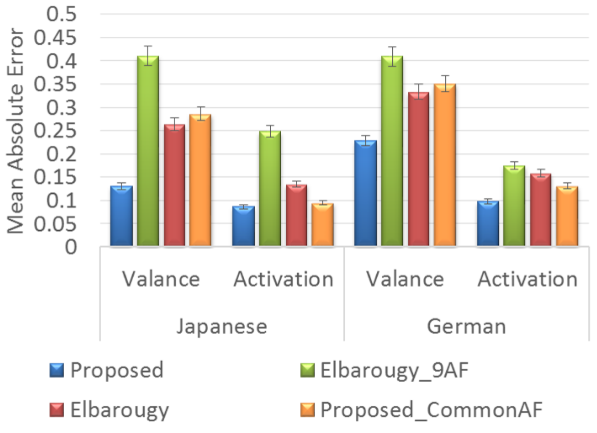


Fig. 4. Mean absolute error between human evaluation and the estimated values of emotion dimensions

Table 3. Japanese recognition accuracy from Bilingual perceptual model in four cases

Method	Japanese Classification rate (%)				
	Neutral	Joy	Anger	Sad	Average
Proposed	95.0%	95.0%	95.0%	100.0%	96.0%
Elbarougy_9AF	90.0%	80.0%	83.0%	80.0%	83.0%
Elbarougy	80.0%	95.0%	90.0%	90.0%	89.0%
Proposed_CommonAF	90.0%	78.0%	70.0%	95.0%	83.0%

Table 4. German recognition accuracy from Bilingual perceptual model in four cases

Method	German Classification rate (%)				
	Neutral	Joy	Anger	Sad	Average
Proposed	92.0%	84.0%	72.0%	90.0%	85.0%
Elbarougy_9AF	86.0%	58.0%	54.0%	100.0%	75.0%
Elbarougy	84.0%	68.0%	66.0%	94.0%	78.0%
Proposed_CommonAF	86.0%	60.0%	62.0%	94.0%	76.0%

The results in the above tables indicate that the proposed system works well for improving bilingual speech emotion dimensions estimation compared with previous study in[12].

4.3. Discussion

As shown in Fig. 4, the best performance for emotion dimensions estimation were achieved using the proposed system, for each emotion dimension, with the smallest MAEs and highest recognition accuracy for both Japanese and German databases. In the case of using nine acoustic features proposed in this paper, compared with Elbarougy_9AF the average MAE of valence and activation of the proposed system was decreased from 0.33 to 0.10 and from 0.29 to 0.16 for Japanese and German respectively. While using the common features that used Elbarougy's study, compared with Elbarougy's system the average MAE of valence and activation of Proposed_CommonAF was reduced from 0.20 to 0.19 and from 0.25 to 0.24 for Japanese and German respectively. For the German and Japanese databases, the overall best results were achieved for all emotion dimensions using the proposed method. Furthermore, The emotion classification accuracies listed in Table 3 and Table 4 just correspond to the MAEs for each systems.

From this discussion, it is evident that estimation of bilingual emotion estimation effectively being improved by using the proposed model. Therefore, the most important results from this study is that the proposed bilingual speech emotion recognition system based on the newly revised three-layered model using a normalization method on emotion dimensional space overcomes the limitation in previous study.

5. CONCLUSION

Using the proposed acoustic features selection method and the novel normalization method in this paper, the estimation of emotion dimensions on bilingual emotion perceptual model have been effectively improved compared with previous work. The most worthy result is that, the proposed method in this paper is suitable to be extended for estimating emotional states conveyed in the source languages in a S2ST system regardless of the languages used in the future work.

6. REFERENCES

[1] M. Akagi, X. Han, and R. Elbarougy, "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages," in *APSIPA*. IEEE, 2014, pp. 1–10.

[2] Björn Schuller, Stefan Steidl, and Anton Batliner, "The interspeech 2009 emotion challenge.," in *INTERSPEECH*, 2009, vol. 2009, pp. 312–315.

[3] R. Elbarougy and M. Akagi, "Improving speech emotion dimensions estimation using a three-layer model of human perception," *Acoustical Science and Technology*, vol. 35, no. 2, pp. 86–98, 2014.

[4] Erickson D. Huang, C. F. and M. Akagi, "Comparison of japanese expressive speech perception by japanese and taiwanese listeners," *Acoustics2008, Paris,2317-2322*, 2008.

[5] Oudeyer Pierre, Y., "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 157–183, 2003.

[6] M. Grimm and K. Kroschel, *Emotion estimation in speech using a 3d emotion space concept*, Citeseer, 2007.

[7] M. Valstar and B. Schuller et.al, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. ACM, 2013, pp. 3–10.

[8] D. Wu and Thomas D Parsons, "Acoustic feature analysis in speech emotion primitives estimation.," in *INTERSPEECH*, 2010, pp. 785–788.

[9] K. R Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487, 1978.

[10] Egon Brunswik, "Historical and thematic relations of psychology to other sciences," *The Scientific Monthly*, vol. 83, pp. 151–161, 1956.

[11] C.F. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.

[12] R. Elbarougy and M. Akagi, "Cross-lingual speech emotion recognition system based on a three-layer model for human perception," Proc. APSIPA2013, Kaohsiung, Taiwan (2013).

[13] X. Han, R. Elbarougy, M. Akagi, J. Li, T. D. Ngo, and T. D. Bui, "A study on perception of emotional states in multiple languages on valence-activation approach," Proc NCSP2015, Kuala Lumpur, Malaysia (2015).

[14] A. Burkhardt, F.and Paeschke, "A database of german emotional speech.," in *Interspeech*, 2005, vol. 5, pp. 1517–1520.

[15] J. Jang, "Anfis: adaptive-network-based fuzzy inference system," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 23, no. 3, pp. 665–685, 1993.