ICSV 21
13-17 July, 2014, Beijing/China

The 21st International Congress on Sound and Vibration

13-17 July, 2014, Beijing/China

TOWARD RELAYING EMOTIONAL STATE FOR SPEECH-TO-SPEECH TRANSLATOR: ESTIMATION OF EMOTIONAL STATE FOR SYNTHESIZING SPEECH WITH EMOTION

Masato Akagi[1] and Reda Elbarougy[1,2]

[1] *Japan Advanced Institute of Science and Technology (JAIST), Japan,*
[2] *Department of Mathematics, Faculty of Science, Damietta University, New Damietta, Egypt*
e-mail: akagi@jaist.ac.jp

Most of the previous studies on Speech-to-Speech Translation (S2ST) focused on processing of linguistic content by directly translating the spoken utterance from the source language to the target language without taking into account the paralinguistic and non-linguistic information like emotional states emitted by the source. However, for clear communication, it is important to capture and transmit the emotional states from the source language to the target language. In order to synthesize the target speech with the emotional state conveyed at the source, a speech emotion recognition system is required to detect the emotional state of the source language. The S2ST system should enable the source and target languages to be used interchangeably, i.e. it should possess the ability to detect the emotional state of the source regardless of the language used. This paper proposes a Bilingual Speech Emotion Recognition (BSER) system for detecting the emotional state of the source language in the S2ST system. In natural speech, humans can detect the emotional states from the speech regardless of the language used. Therefore, this study demonstrates feasibility of constructing a global BSER system that has the ability to recognize universal emotions. This paper introduces a three-layer model: emotion dimensions in the top layer, semantic primitives in the middle layer, and acoustic features in the bottom layer. The experimental results reveal that the proposed system precisely estimates the emotion dimensions cross-lingual working with Japanese and German languages. The most important outcome is that, using the proposed normalization method for acoustic features, we found that emotion recognition is language independent. Therefore, this system can be extended for estimating the emotional state conveyed in the source languages in a S2ST system for several language pairs.

1. Introduction

Speech-to-Speech Translation (S2ST) is the process by which a spoken utterance in one language is used to produce a spoken output in another language. Traditionally automatic speech translation consists of three component technologies: converting the spoken utterance into a text using an Automatic Speech Recognition (ASR) system, then the recognized speech is translated using a Machine Translation (MT) system into the target language text, finally, resynthesizes the target language text using a text-to-speech (TTS) synthesizer [1, 2]. Therefor, the traditional approach for

This paper is based on one first published in the proceedings of the 21st International Congress on Sound and Vibration, July 2014 and is published here with permission of the International Institute of Acoustics and Vibration (IIAV.)

S2ST focused on processing of linguistic content only by directly translating the spoken utterance from the source language to the target language without taking into account the paralinguistic and non-linguistic information like emotional states emitted by the source. Conventional S2ST systems output speech usually produced in a neutral voice that is unchanged even if the input speech changes from emotional state to another. However, for a natural communication, it is important to preserve the emotional states expressed in the source language.

This study investigates how to transform the emotional states from the source language to the target language in a S2ST system. Therefor, in order to accomplish this task, two additional components are necessary. The first is an automatic emotion recognition system to detect the emotional state of the source speech. Moreover, the second is an emotional speech synthesizer to synthesize the target speech with the emotional state conveyed at the source.

In this paper, we focused on the first component that is constructing an automatic speech emotion recognition system that has the ability to detect the emotional state in the source language. The S2ST system should enable the source and target languages to be used interchangeably, i.e. it should possess the ability to detect the emotional state of the source regardless of the language used. Changing the source language require changing the training language for the speech emotion recognition system, i.e. adapting the system for different language. However, human can still judge the expressive content of a voice for one language, such as emotional states, even without the understanding of that language [3]. Several studies have indeed shown evidence for certain universal attributes for speech [4, 5], not only among individuals of the same culture, but also across cultures. Therefore, in order to overcome the problem of retrain the emotion recognition system for different language, this paper proposes a Bilingual Speech Emotion Recognition (BSER) system for detecting the emotional state of the source language in the S2ST system.

In order to produce the output of the S2ST system colored with emotional state of the speaker at the source, firstly, it is required to detect the emotional state at the source, then, modifying the acoustic features of the neutral speech produced by TTS system to an emotional speech. In the literature the emotional states can be represented by the categorical approach such as happy, anger or can be represented as a point in n-dimensional space, such as the valence-activation-dominance space [6, 7]. The categorical and dimensional approaches are closely related, i.e. by detecting the emotional content using one of these two schemes, we can infer its equivalents in the other scheme. Emotional space representation is more convenient for the task of emotion recognition for the following reasons: (1) Using this method, we can determine not only the emotion category but also the degree of that state for example very happy, little happy and so on. (2) In most of the previous studies the emotional space are very similar i.e. in most of the happy state is exist in the first quarter of the valence-activation space. (3) Any improvement in the dimensional approach will lead to an improvement in the categorical approach and vice versa [8].

In this study, we assume that the emotional spaces are identical for different languages i.e. the distance and directions from neutral voice to other emotional states are common among languages. In order to re-synthesize the target language with emotional state by modifying the acoustic features from neutral to an emotional state, the following steps are required:

- extract variety of acoustic features of the source language, and selecting the most related features to emotion dimensions.
- estimate the emotion dimensions valence activation and dominance, using a speech emotion recognition system.
- investigates whether the acoustic features to realize specific emotions are language independent.
- finding the acoustic features corresponding to the estimated emotion dimensions in step 2.

This paper is based on one first published in the proceedings of the 21st International Congress on Sound and Vibration, July 2014 and is published here with permission of the International Institute of Acoustics and Vibration (IIAV.)

- modifying the acoustic feature of the neutral speech produced by TTS system to the acoustic features determined in the previous step.

This study focused on the estimation of the emotional state of the source language and investigating if the acoustic features realization of specific emotions is language independent. The findings can guide us to modify the emotional state of the target language to preserve the emotional state of the source language. Elbarougy and Akagi proposed a three-layer model for estimating emotion dimensions: valence, activation, and dominance [9]. The prediction accuracy for estimating emotion dimensions was improved using this model [10]. Therefore, in this study, the proposed three-layer model is used to detect the emotional state of the source language in a S2ST system. In order to evaluate the proposed system, the estimated values of emotion dimensions (valence, activation, and dominance) were mapped to emotion categories using Gaussian Mixture Model (GMM). The classifications results of the proposed system were compared with the results of the traditional categorical approach, which map acoustic features directly to emotion categories. To investigate whether acoustic features realization of specific emotions is language independent, the classification results into emotion categories were compared in three different cases: mono-language, cross-language, and bilingual emotion recognition.

2. Automatic Emotion Recognition System

This section investigates the design of a bilingual emotion recognition system based on the three-layer model. The emotional states in this paper are represented by the dimensional approach. This approach defines emotions as points in a three-dimensional emotion space spanned by the three basic dimensions valence (negative-positive axis), activation (calm-excited axis), and dominance (weak-strong axis). The task of emotion recognition using the dimensional approach can be viewed as using an estimator to map the acoustic features to real-valued emotion dimensions. Every point in the dimensional space can be mapped into one emotion category. The block diagram of the proposed method for emotion dimension estimations is shown in Fig. 1. The extracted acoustic features are mapped into emotion dimensions, valence, activation and dominance, using the three-layer model.

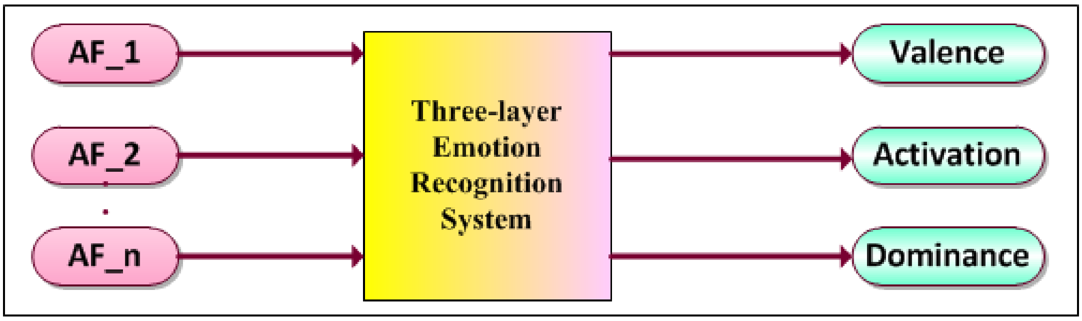


Figure 1. Block diagram of the proposed approach for emotion dimensions estimation.

In order to construct the proposed bilingual emotion recognition system, at least two databases in different languages are requited. The elements of the proposed emotion recognition system were collected in following section.

2.1 Speech Material

In order to validate the proposed method two emotional speech databases were used. One of these databases is in the Japanese language and the other is in the German language. The Japanese database is the multi-emotion single speaker Fujitsu database produced and recorded by Fujitsu Laboratory, it contains five emotional states: neutral, joy, cold anger, sad, and hot anger as described in [9].

This paper is based on one first published in the proceedings of the 21st International Congress on Sound and Vibration, July 2014 and is published here with permission of the International Institute of Acoustics and Vibration (IIAV.)

The German database is the Berlin database [11]. It comprises seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral speech. An equal distribution of the four similar emotional states (neutral, happy, angry, and sad) as follows: 50 happy, 50 angry, 50 sad, and 50 neutral, totally 200 utterances were selected from the Berlin database.

2.2 The elements of the proposed model

In this study, the human perception model as described by Scherer [12] is adopted. This model assumes that human perception is a multi-layer process. It was assumed that the acoustic features are perceived by a listener and internally represented by a smaller perception e.g., adjectives describing emotional voice as reported by Huang and Akagi [13]. These smaller percepts or adjectives are finally used to detect the emotional state of the speaker. Human subjects can subjectively evaluate these adjectives using a listening test. Therefore, the following set of adjectives describing the emotional speech were selected as candidates for semantic primitives: bright, dark, high, low, strong, weak, calm, unstable, well-modulated, monotonous, heavy, clear, noisy, quiet, sharp, fast, and slow. The proposed model consists of three layers: emotion dimensions valence, activation and dominance, which constitute the top layer, semantic primitives that constitute the middle layer and acoustic features that form the bottom layer. The semantic primitive layer is added between the traditional layers (emotion dimensions and acoustic features) to imitate human perception as described by Scherer [12], in his description of human perception adopted a version of Brunswik's lens model which was originally proposed in 1956 [14].

For constructing a speech emotion recognition system based on the proposed model, many acoustic features must be extracted, semantic primitives and the three emotion dimensions must be evaluated for each utterance in the two databases. Therefore, two listening test was used to evaluate semantic primitives and emotion dimensions as explained in [9], for the two databases. Moreover, an initial set of 21 acoustic features were extracted for each utterance in the two databases. In order to avoid speaker and language dependency on the acoustic features, acoustic feature normalization is done, in which all-acoustic feature values are normalized by those of the neutral speech. This was performed by dividing the values of acoustic features by the mean value of neutral utterances for all acoustic features. Then, the feature selection method proposed by Elbarougy and Akagi was used to select the most related acoustic features for each emotion dimensions [10].

2.3 System Implementation

Having identified the best acoustic features set, we constructed an individual estimator to predict the values (-1 to 1 rated by the listening test) of each emotion dimension based on the three-layer model. The three-layer model imitates human perception by estimating the adjectives describing the emotional speech, followed by estimating emotion dimensions from the estimated adjectives.

In order to implement the proposed system an Adaptive-Network-based Fuzzy Inference System (ANFIS) was used to connect the elements of this system. For example, in order to estimate the valence dimension using the proposed model as described in Fig. 2. This figure shows the elements of the three-layer used for estimation. These layers are as follows: the first layer consists of 6 acoustic features; MH_A, MH_E and MH_O are the mean value of H1-H2 for vowels /a/, /e/, and /o/, respectively, F0_RS is the rising slope of F0 contour, F0_HP highest F0, and power range PW_R. The second layer is the most related semantic primitives or the most related adjectives describing the valence dimensions. A bottom-up method was used to estimate the values (1 to 5 rated by the listening test) of the six semantic primitives in the middle layer from the six acoustic features in the bottom layer. Since, FIS has the structure of multiple inputs and of one output [15]. Therefore, in order to estimate valence, seven FISs are needed, six FISs to estimate each semantic primitive, in addition, one to estimate the value of the valence dimension from the estimated six semantic primitives. In a similar way, the activation and dominance can be estimated using FIS for each semantic primitive,

This paper is based on one first published in the proceedings of the 21st International Congress on Sound and Vibration, July 2014 and is published here with permission of the International Institute of Acoustics and Vibration (IIAV.)

and one FIS for the activation and dominance respectively. The next section describes the evaluation of the proposed system.

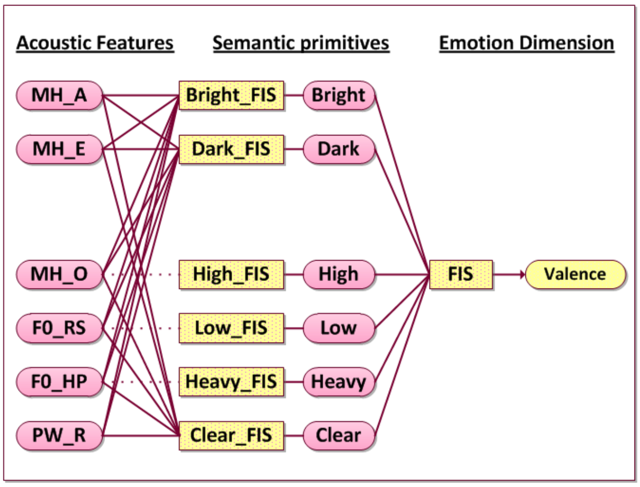


Figure 2. Block diagram of the proposed approach for estimating valence based on the three-layer model.

3. System Evaluation

The aim of this study is estimate the emotional state of the source language of S2ST system. Therefore, in the previous section we introduce an emotion recognition system to estimate emotion dimensions valence, activation and dominance for the source language. This section investigates whether the proposed system trained using bi-languages has the ability to detect the emotion dimension for different languages as explained in Subsection 3.1. In order to investigate the improvement of emotion categorical classification, the estimated values of emotion dimensions were used as an input features to train GMM classifier to classify emotional state into emotion categories as explained in Subsection 3.2.

3.1 Results for emotion dimension estimations

For evaluation of the proposed system, the Mean Absolute Error (MAE) and the correlation were used to compare between the human evaluations using the listening test ant the system output. The mean absolute error MAE between the predicted values of emotion dimensions and the corresponding average value given in listening tests by human subjects is used as a metric of the discrimination associated with each case. The MAE is calculated according to the following equation:

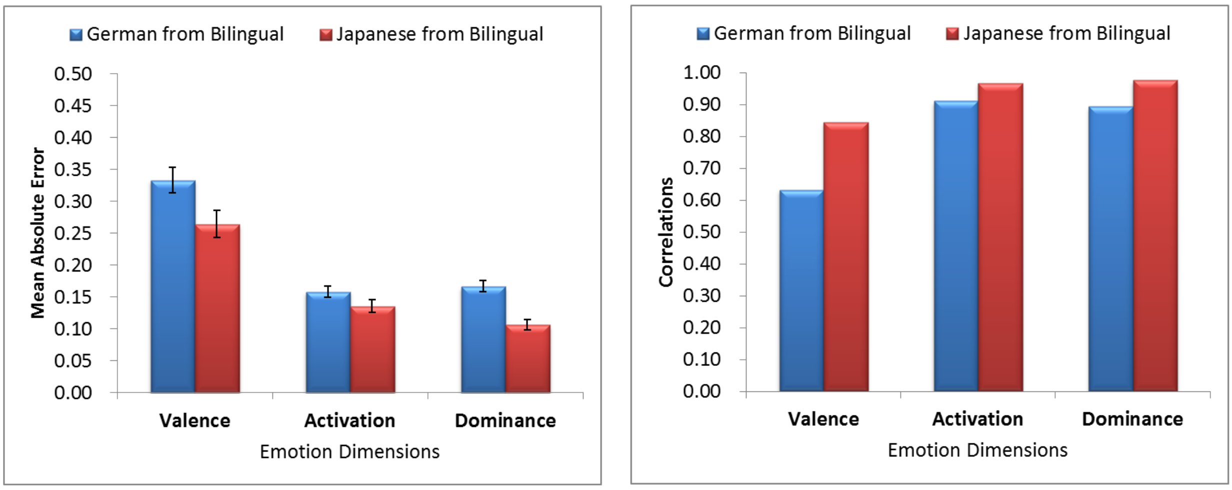
$$MAE^{(j)} = \frac{\sum_{i=1}^{N} |\widehat{x}_i^{(j)} - x_i^{(j)}|}{N} \tag{1}$$

where $j \in \{valence, activation, dominance\}$, $\widehat{x}_i^{(j)}$ is output of the emotion recognition system, and $x_i^{(j)}$, $-1 \leq x_i^{(j)} \leq 1$ is the values evaluated by the human subjects in listening tests.

The proposed system was trained using the combined information for acoustic feature information, semantic primitive, emotion dimensions for both languages, which called bilingual emotion dimensions estimation. In order to estimate emotion dimensions for Japanese utterances the acoustic features for that utterance is used as an input to the trained bilingual system. The MAEs for all emotion dimensions, for both Japanese and German database were presented in Fig. 3(a). From this figure, the MAE for estimating Japanese emotion dimensions from the bilingual system is as follows: 0.26, 0.14 and 0.11 for valence, activation and dominance, receptively. While, for German database the estimation results were 0.33, 0.16 and 0.17. These results reveal that the estimation for activation

This paper is based on one first published in the proceedings of the 21st International Congress on Sound and Vibration, July 2014 and is published here with permission of the International Institute of Acoustics and Vibration (IIAV.)

and dominance are very close to human evaluation for both databases. However, there are small error for estimating valence for valence for both databases, however, these error do not constitute a large difference comparing to the used range for evaluation from -1 to 1. The correlations between the estimated values of emotion dimensions and the evaluated values using the listening tests are presented as shown in Fig. 3(b). The values of the correlations for activation and dominance were very high which indicate the best estimation accuracy for them in both languages.



(a) Mean absolute error between human evaluation and the estimated values of emotion dimensions

(b) Correlation between human evaluation and the estimated values of emotion dimensions

Figure 3. Estimation accuracy of the first subsystem for emotion dimension estimation for Japanese and German database using bilingual estimation system.

These results ravel that the proposed method accurately estimate emotion dimensions for both languages Japanese and German. Therefore, this indicates that the proposed method can detect the emotional state regardless of the used language at the source of the S2ST system.

3.2 Results for emotion classification

Every point in the emotional space can be mapped into emotion categories. Therefore, this section evaluates the corresponding categorical classification to the estimated emotional space using the proposed method. GMM classifier was used to map the estimated emotion dimensions into emotion categories. This section also investigates whether the acoustic feature realization of specific emotion is language independent.

In order to evaluate the categorical classification, the results of the proposed system were compared with those of the traditional categorical method that map acoustic features directly to emotion category using GMM classier. For investigating the language independent for emotion classification, the performance of the proposed bilingual system, were compared with those of the mono-language and cross-language emotion recognition system. In case of mono-language, the system is trained and is tested using the same language, and in case of cross-language, the system is trained using one language and tested using the second language. The results of the traditional and the proposed system are shown in Tables 1 and 2 for Japanese language, respectively, and in Tables 3 and 4 for German language.

The emotion classification accuracies listed in Tables 1- 4 correspond to the MAEs for the estimated emotion dimensions using the proposed system described in the above section. It is clearly seen that the recognition rate using the proposed method outperforms the results using the traditional categorical approach. Comparing the classification results for the bilingual system with the mono language system it was found that the difference is small for German language recognition rate decreased from 75.0% to 68.7%, which is not so large difference. For Japanese language, the results

This paper is based on one first published in the proceedings of the 21st International Congress on Sound and Vibration, July 2014 and is published here with permission of the International Institute of Acoustics and Vibration (IIAV.)

Table 1. Classification rate for Japanese using the traditional approach by mapping acoustic features into emotion categories using GMM classifier, in following cases: (1) mono-language, (2) cross-language and (3) bilingual emotion recognition system.

The used method	Classification rate (%)				
	Neutral	Joy	Sad	Hot Anger	Average
Japanese from Japanese	68.8	46.9	100.0	65.6	**70.3**
Japanese from German	75.0	81.3	68.8	9.4	**58.6**
Japanese from Bilingual	75.0	46.9	100.0	65.6	**71.9**

Table 2. Classification rate for Japanese using the proposed approach by mapping the estimated values of emotion dimensions using the three-layer model into emotion categories using GMM classifier, in following cases: (1) mono-language, (2) cross-language and (3) bilingual emotion recognition system.

The used method	Classification rate (%)				
	Neutral	Joy	Sad	Hot Anger	Average
Japanese from Japanese	80.0	97.5	100.0	92.5	**92.5**
Japanese from German	95.0	100.0	100.0	70.0	**91.3**
Japanese from Bilingual	75.0	84.4	100.0	90.6	**87.5**

Table 3. Classification rate for German using the traditional approach by mapping acoustic features into emotion categories using GMM classifier, in following cases: (1) mono-language, (2) cross-language and (3) bilingual emotion recognition system.

The used method	Classification rate (%)				
	Neutral	Happy	Sad	Anger	Average
German from German	57.5	42.5	80.0	62.5	**60.6**
German from Japanese	22.5	50.0	40.0	42.5	**38.8**
German from Bilingual	60.0	62.0	77.5	42.5	**60.5**

Table 4. Classification rate for German using the proposed approach by mapping the estimated values of emotion dimensions using the three-layer model into emotion categories using GMM classifier, in following cases: (1) mono-language, (2) cross-language and (3) bilingual emotion recognition system.

The used method	Classification rate (%)				
	Neutral	Happy	Sad	Anger	Average
German from German	74.0	62.0	80.0	84.0	**75.0**
German from Japanese	40.0	87.5	72.5	42.5	**60.6**
German from Bilingual	75.0	67.5	62.2	70.0	**68.7**

decreased from 92.5% to 87.5% using the proposed method, which is very small error. These results indicate that bilingual emotion recognition system can be used to classify the emotional state for both languages with a small error. Therefore, this method improves the classification rate for both languages. The classification results using the proposed method as shown in Table 2 and 4 indicate a small difference between the mono-language, cross-language and the bilingual cases, which reveal that the acoustic feature realization is language independent.

4. Conclusion

In this paper, we attempted to construct a general emotion recognition system in order to detect the emotional states of the source language in a speech-to-speech translation system. Therefore, this study proposed a bilingual emotion recognition system based on a three-layer model of human perception. The proposed system was trained using the combined information form two different languages (Japanese and German). In order to estimate emotion dimensions for both languages, the acoustic features are used as input to the trained bilingual emotion recognition system. The estimation results

This paper is based on one first published in the proceedings of the 21st International Congress on Sound and Vibration, July 2014 and is published here with permission of the International Institute of Acoustics and Vibration (IIAV.)

for emotion dimensions reveal that the proposed system effectively estimate emotion dimensions for the two languages.

The GMM classifier was used to map the estimated emotion dimensions into emotion categories. By comparing the classification results for the proposed bilingual system and the mono-language system, it was found that a small difference for both languages 5.0%, 6.3% for Japanese and German language, respectively. These results indicate that bilingual emotion recognition system can be used effectively to detect the emotional state for the source language in a S2ST system regardless of the used language.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Nakamura, S. Overcoming the language barrier with speech translation technology, *NISTEP Quarterly Review*, **31**, pp. 35–48, (2009).

[2] Shimizu, T., Ashikari, Y., Sumita, E., Zhang, J.S. and Nakamura, S. NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System, *Tsinghua Science and Technology*, **13**(4), pp. 540–544, Aug. (2008).

[3] Huang, C. F., Erickson, D., and Akagi, M. Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners, *Acoustics2008*, Paris, pp. 2317–2322, (2008).

[4] Banse, R., and Scherer, K.R. Acoustic profiles in vocal emotion expression, *Journal of personality and social psychology*, **70**(3), pp. 614–636, March (1996).

[5] Scherer, K.R., Banse, R., Wallbott, H.G., and Goldbeck, T. Vocal cues in emotion encoding and decoding. Motivation and Emotion, **15**(2), pp. 123–148, June (1991).

[6] Lee, C.M., and Narayanan, S. Toward Detecting Emotions in Spoken Dialogs, *IEEE Transactions on Speech and Audio Processing*, **13**(2), pp. 293–303, (2005).

[7] Wu, D., Parsons, T.D., and Narayanan, S. Acoustic Feature Analysis in Speech Emotion Primitives Estimation, *Proc. InterSpeech 2010*, pp. 785–788, (2010).

[8] Grimm, M., Kroschel, K., Mower, E., and Narayanan, S. Primitives-based evaluation and estimation of emotions in speech, *Speech Communication*, **49**, pp. 787–800, (2007).

[9] Elbarougy, R., and Akagi, M. Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model, *Proc. Int. Conf. APSIPA ASC*, (2012).

[10] Elbarougy, R., and Akagi, M. Improving speech emotion dimensions estimation using a three-layer model of human perception, *Acoust. Sci & Tech.*, **35**(2), (2014).

[11] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. A Database of German Emotional Speech, *Proceedings of Interspeech*, Lissabon, Portugal, (2005).

[12] Scherer, K.R. Personality inference from voice quality: The loud voice of extroversion, *European Journal of Social Psychology*, **8**, pp. 467–487, (1978).

[13] Huang, C., and Akagi, M. A three-layered model for expressive speech perception, *Speech Communication*, **50**(10), pp. 810–828, October (2008).

[14] Brunswik, E. Historical and thematic relations of psychology to other sciences, *Scientific Monthly*, **83**, pp. 151–161, (1956).

[15] Jang, J.-S.R. ANFIS: Adaptive network-based fuzzy inference system, *IEEE Transactions on Systems, Man and Cybernetics*, **23**(3), pp. 665–685, (1993).

This paper is based on one first published in the proceedings of the 21st International Congress on Sound and Vibration, July 2014 and is published here with permission of the International Institute of Acoustics and Vibration (IIAV.)