TOWARD RELAYING AN AFFECTIVE SPEECH-TO-SPEECH TRANSLATOR: CROSS-LANGUAGE PERCEPTION OF EMOTIONAL STATE REPRESENTED BY EMOTION DIMENSIONS

Reda Elbarougy [1,2]*, Han Xiao* [1]*, Masato Akagi* [1]*, and Junfeng Li* [3]

[1] Japan Advanced Institute of Science and Technology, Japan
[2] Dep. of Math., Faculty of Science, Damietta University, Egypt
[3] Institute of Acoustics, Chinese Academy of Science, China

elbarougy@jaist.ac.jp, han_xiao@jaist.ac.jp, akagi@jaist.ac.jp, lijunfeng@hccl.ioa.ac.cn

ABSTRACT

Affective speech-to-speech translation (S2ST) is to preserve the affective state conveyed in the speaker's message. The ultimate goal of this study is to construct an affective S2ST system that has the ability to transform the emotional states of a spoken utterance from one language to another language. A universal automatic speech-emotion-recognition system is required to detect emotional state regardless of language. Therefore, this study investigates commonalities and differences of emotion perception across multi-languages. Thirty subjects from three countries, Japan, China and Vietnam, evaluate three emotional speech databases, Japanese, Chinese and German, in valence-activation space. The results reveal that directions from neutral to other emotions are similar among subjects groups. However, the estimated degree of emotional state depend on the expressed emotional styles. Moreover, neutral positions were significantly different among subjects groups. Thus, directions and distances from neutral to other emotions could be adopted as features to recognize emotional states for multi-languages.

Index Terms— Spoken human-human interaction with multi languages, commonalities of emotion perception in multi-languages, human perception.

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. The conventional automatic speech translation consists of three components: converting the spoken utterance into a text using an automatic speech recognition system, then the recognized speech is translated using a machine translation system into the target language text, and finally, resynthesizes the target language text using a text-to-speech (TTS) synthesizer [1, 2]. Therefore, the traditional approach for S2ST focused on processing of linguistic content only by 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 neutral voice, which is unchanged even if the input speech changes from one emotional state to another. However, for a natural communication, it is important to preserve the emotional states expressed in the source language [3].

To produce an output of the S2ST system colored with emotional states of the speakers at the source, firstly, it is required to detect the emotional state at the source, then, modify the acoustic features of the neutral speech produced by TTS system to emotional speech. Changing the source language requires changing the training language for the Speech Emotion Recognition System (SERS), i.e. adapting the system a for different language. Therefore, the required SERS 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 language [4, 5].

People can still judge the expressive content of a voice for one language, such as emotional states, even without the understanding of that language [6]. Several studies have indeed shown evidence for certain universal attributes for speech [7], not only among individuals of the same culture, but also across cultures. Therefore, the aim of this study is to investigate commonalities of human perception for emotional speech among multi-languages. These investigations can help us to acquire fundamental knowledge on how human beings perceive emotional states of different languages. Understanding human perception for multi-languages will be able to guide us to construct a global automatic emotion recognition system.

2. STRATEGY OF THE STUDY

In the literature, the emotional states are represented as categorical items such as happy, anger, or are represented as a point in n-dimensional space, such as Valence-Activation (V-

A) space [8]. Emotion categories can be represented by regions in the V-A space, where the neutral state lies in the center, and other emotions lie in a specific region in this space [9]. For example, happy is represented by a region in the first quarter, in which valence is positive, and activation is high.

Several studies compared human perception for emotional speech among multi-languages using the categorical representation, although, using this representation for investigating human perception will not reflect the similarity and differences among different languages. However, the dimensional representation using V-A space is more convenient for the task of comparing human perception among different languages for the following reasons. (1) In V-A space it is easy to compare the relative position of emotional states, for instance, to compare neutral positions among different languages. This information can be used to normalize acoustic features in order to adapt multi-languages. (2) To determine the direction from neutral to other emotional states which can help us find the common directional acoustic features among different languages.

3. EXPERIMENT

In this Section, the commonalities and differences in human perception of emotional speech among multi-languages will be investigated in the V-A space. Therefore, the values of valence and activation are evaluated for three emotional speech databases using three different languages: Japanese, German and Chinese. These language were selected because it is usually not the second language for many people. Thirty subjects from three different countries: Japan, China and Vietnam evaluated the emotional contents using listening test.

Four emotional states: happy, angry, neutral, and sad were selected from the three databases. The central positions for these emotions are calculated by the average value of valence and activation for each emotion category separately. The central positions of all emotional states are compared for the three listener groups, for each database individually. This comparison can guide us to find the commonalities and differences among languages. Moreover, the differences between languages can be used to normalize acoustic features extracted from multi-languages in order to construct a universal automatic SERS for the recognition part of an S2ST system.

3.1. Database

Three emotional speech databases were selected for conducting this study, Japanese, German, and Chinese, and all of them consisted of acted emotions. In order to compare the results among the three databases only similar categories were selected. Therefore, the four similar categories are neutral, happy, anger and sad. The Japanese database is the Fujitsu database produced and recorded by Fujitsu Laboratory. In this database, a professional actress was asked to produce utterances using 5 emotional categories, i.e. neutral, happy, cold anger, sadness, and hot anger. There are 20 different sentences, each sentence has one speech utterance in neutral and two speech utterances in each of the other categories. Thus, for each sentence, there are 9 utterances and for all 20 sentences, there are 180 utterances. Therefore, for the Japanese database the selected utterances were as follows: 20 neutral, 40 happy, 40 anger and 40 sad, a total of 140 utterances.

The German database is the Berlin database, which comprises seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral [10]. Ten professional German actors (five females and five males) spoke ten sentences with an emotionally neutral content in seven different emotions. These sentences were not equally distributed between the various emotional states: 127 angry, 81 bored, 46 disgusted, 69 frightened, 71 happy, 61 sad, 79 neutral. Finally, in total, 200 utterances were selected from the Berlin database, 50 utterances of four similar emotional states as the Japanese database.

The Chinese database of emotional speech was produced and recorded by CASIA by 2 professional actresses and 2 professional actors. For each speaker, there are 6 different emotions: neutral, happy, anger, sad, fear and surprise. The four speakers spoke 400 sentences once using one emotional state. Therefore, 192 utterances were selected from the Chinese database as follows: 48 neutral, 48 happy, 48 anger and 48 sad.

3.2. Subjects

The three databases were evaluated in terms of valence and activation using listening tests with thirty subjects: 10 Japanese, 10 Chinese, and 10 Vietnamese. All subjects are graduate students, 5 subjects of each nationality were males and the other 5 were females. Subjects ages are from 20 to 35. No subjects have hearing impairments.

Japanese and Chinese subjects were selected to compare the evaluation results of native speakers with non-native speaker. However, it was difficult to find German speakers to participate as subjects. Therefore, Vietnamese subjects were selected to investigate whether subjects can recognize the emotional state without understanding all the used languages. No subjects have any knowledge about the German language. Japanese and Vietnamese subjects do not have any knowledge about the Chinese language.

3.3. Procedure

Most of the existing emotional speech databases were annotated using the categorical approach, including the selected databases. Therefore, listening tests are required to annotate each utterance in the used databases using the dimensional representation. Thus, the three emotional databases were evaluated through listening tests in terms of valence and

activation. Each emotional speech database was evaluated three times by the three listener groups: Japanese, Chinese, and Vietnamese. For valence and activation evaluations, a 5-point scale {-2, -1, 0, 1, 2} was used. Valence scale is very negative (-2), negative (-1), neutral (0), positive (1), and very positive (2). Activation scale is very calm (-2), calm (-1), neutral (0), excited (1), and very excited (2). Before starting the experiment the basic theory of dimensional representation was explained for subjects following the work of Mori et al. [11]. Then they took a training session to listen to an example set composed of 16 utterances, which covered the five-point scale. The purpose of this training set is to let subjects understand the dimensional representation. A pre-test was conducted to test the subjects' ability to score the emotional sensitivity. In this test, subjects were asked to evaluate another set of 16 utterances different from those used in the training. Subjects who gained more than 60% accuracy in the pre-test were allowed to participate in the main-test.

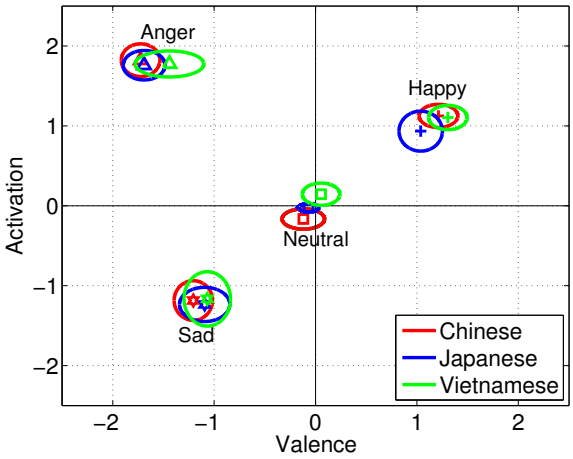
In the main-test, a MATLAB GUI was used for evaluation. All stimuli from the three databases were presented randomly through binaural headphones at a comfortable sound pressure level in a soundproof room. Subjects were asked to evaluate their perceived impression from the way of speaking, not from the content itself, and then choose a score on the five-point scale for valence and activation individually in two different sessions. The average of the subjects' rating for valence and activation was calculated per utterance.

3.4. Agreement between subjects

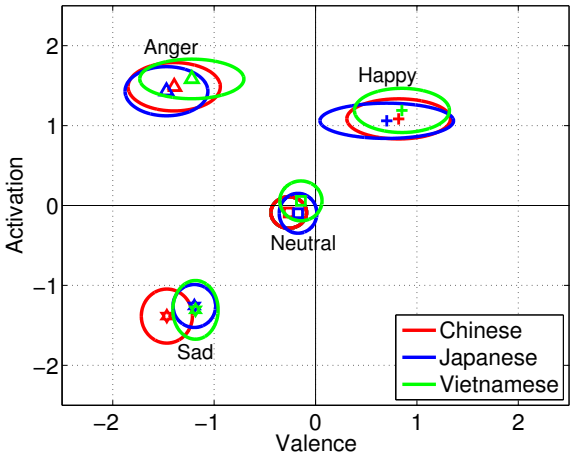
The inter-rater agreement was measured by means of pairwise Pearson's correlations between every two subjects' ratings, separately for valence and activation. Subjects who gained lower than 0.60 correlations with other subjects were removed from the final analysis. Finally, it was found that all subjects agreed from moderate to a high degree for valence and activation.

4. EXPERIMENTAL RESULTS

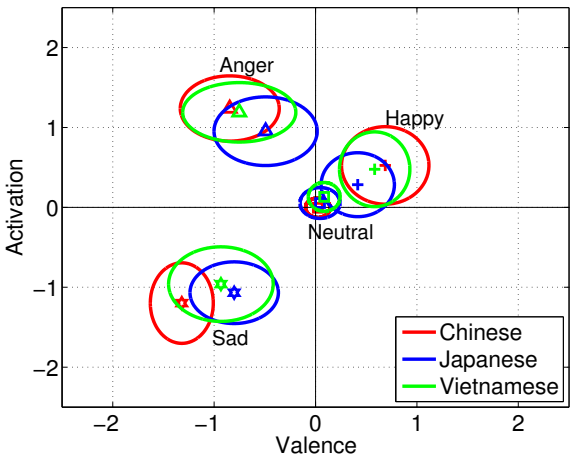
The results of the listening experiments for the three subject groups in the V-A space were compared as follows. Each emotional state (E) represented by ellipse in the V-A space, where the center of this ellipse is (x_E, y_E) and where x_E and y_E are the average of valence and activation of the emotional state E. Moreover, the horizontal and vertical radius of this ellipse are the standard deviation of valence and activation, respectively. The center of each ellipse was considered the central position of the represented emotional state. Figure 1 shows the position of emotional states: happy, angry, neutral and sad in V-A space for the three databases: (a) Japanese, (b) German, and (c) Chinese. The colors Red, Blue and Green represent the three subjects groups Chinese, Japanese and Vietnamese, respectively.



(a) Japanese Database



(b) German Database



(c) Chinese Database

Fig. 1. Emotional states position in valence-activation space.

4.1. Neutral position

In order to investigate whether the position of neutral states are the same or not for the three listener groups, statistical

analysis was used. Analysis of variance (ANOVA) was conducted to check whether there is a significant difference between the neutral positions among subject groups. The results were as follows: For the Japanese database, a significant difference among the three groups was observed for both valence and activation: valence (F[2, 57] = 5.54 , $p \leq 0.05$), activation (F[2, 57] = 39.52, $p \leq 0.05$). Moreover, the results for the German database were as follows: valence (F[2, 147] = 4.89, $p \leq 0.05$), activation (F[2, 147] = 8.60, $p \leq 0.05$). For the Chinese database there was a significant difference for the position of activation (F[2, 141] = 7.69, $p \leq 0.05$). However, the results for valence were not significant. Therefore, the results for each pair of subjects were compared to check the difference between each pair. It is found that the there is a significant difference between the Chinese and Vietnamese where (F[1, 94] = 4.78, $p \leq 0.05$).

For the Japanese database it is clear from Fig. 1(a) that the Japanese response is in the center of the V-A space but, the Chinese response is negative and calm while the Vietnamese response is positive and exited. For the Chinese database as shown in Fig. 1(c), the Chinese response is in the center, but other listeners responses are positive and strong. For the German database as shown in Fig. 1(b), since German is not the mother language for any of the subjects, all listeners responses are negative; Chinese and Japanese are calm but Vietnamese are exited. These results reveal that the positions of neutral states in three languages are different i.e. the neutral positions depend on the subjects-native language.

4.2. Direction from neutral to emotions

In order to investigate whether the direction from the neutral state to other emotional states are similar or not among listeners, the angle between the direction from neutral to emotional state and the horizontal direction towards positive valence is used as a metric. This angle is calculated by the following equation:

$$angle = arctan(\frac{y_E - y_N}{x_E - x_N}) \quad (1)$$

where (x_E, y_E) is the center of the emotional state E, and (x_N, y_N) is the center of the neutral state N, as explained above. Table 1 lists the angle between the direction from neutral to other emotional states and the horizontal direction towards positive valence for the three subject groups. By comparing these results, it is clear that the direction from neutral to other emotions was similar for the three listener groups, for all databases. For example, in the case of the Japanese database as shown in Table 1(a) the directions from Neutral-to-Anger are very similar: 132°, 129° and 133° for Japanese, Chinese and Vietnamese subjects, respectively. The same trend was observed for all emotions in the German and Chinese databases as shown in detail in Tables 1(a), 1(b) and 1(c). Therefore, we can conclude that directions from neutral state to emotions on V-A space are similar among languages. This result suggests that we can perceive emotional states categorically even not understanding languages. However, from this result it is not clear if human can also perceive the same or different degree of emotions. Therefore, in the next subsection we will investigate whether different listener groups perceive the same degree of emotional state or not.

Table 1. Angles between direction from neutral to other emotions and horizontal direction towards positive valence.

(a) Japanese Database.

Angle (deg.)	Japanese	Chinese	Vietnamese
Neutral-Happy	41°	44°	38°
Neutral-Anger	132°	129°	133°
Neutral-Sad	230°	223°	229°

(b) German Database.

Angle (deg.)	Japanese	Chinese	Vietnamese
Neutral-Happy	53°	47°	49°
Neutral-Anger	130°	126°	125°
Neutral-Sad	229°	227°	233°

(c) Chinese Database.

Angle (deg.)	Japanese	Chinese	Vietnamese
Neutral-Happy	32°	39°	35°
Neutral-Anger	121°	125°	128°
Neutral-Sad	233°	222°	227°

4.3. Degree of emotional state

The distance from the neutral to each emotional state in the V-A space represents the degree of emotional state. The greater the distance is, the stronger the emotion and vice-versa.Therefore, in order to investigate whether the degree of the emotional state is similar or not among different languages, the Euclidean-distance between the neutral state and other emotional states, happy, anger and sad, is used as a metric. The distance between the center of each emotional state and the center of neutral state is calculated using the following equation:

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

The definition of the symbols is the same as in Eq. (1). The results are listed in Table 2. By comparing the degree of emotional states for the three subject groups for each database individually, it was found that the degrees are similar for the three groups for the Japanese and German databases. For the German database, as listed in Table 1(b) the distance between anger and neutral was very similar: 2.00, 1.94 and 1.87 for Japanese, Chinese and Vietnamese subjects, respectively.

However, for the Chinese database there are differences among listeners as follows: (1) the distance between Neutral-Happy which represents the degree of happiness varied from

Table 2. Distance between neutral and other emotional states.

(a) Japanese Database.

Distance	Japanese	Chinese	Vietnamese
Neutral-Happy	1.46	1.85	1.58
Neutral-Anger	2.41	2.55	2.21
Neutral-Sad	1.59	1.49	1.73

(b) German Database.

Distance	Japanese	Chinese	Vietnamese
Neutral-Happy	1.45	1.60	1.51
Neutral-Anger	2.00	1.94	1.87
Neutral-Sad	1.55	1.77	1.72

(c) Chinese Database.

Distance	Japanese	Chinese	Vietnamese
Neutral-Happy	0.44	0.85	0.61
Neutral-Anger	1.05	1.52	1.35
Neutral-Sad	1.41	1.80	1.49

0.44 for Japanese listeners to 0.85 for Chinese listeners, (2) the degree of anger varies from 1.05 for Japanese listeners to 1.52 for Chinese listeners, (3) the degree of sadness ranged from 1.41 to 1.80 for Japanese and Chinese, respectively.

These results suggest that the reason behind these variation is how well each emotional state in each database well performed. The Japanese and German database were recorded using professional actress and actors; therefore, the emotional state was well-performed which indicates why the degrees of emotional stat are similar. However, the Chinese database was recorded by non-professional actors; therefore, the emotional degree varies from speaker to the speaker. These results indicate that subjects using different languages perceive different degrees of emotional state depending on how well performed the emotions were. However, native listeners can perceive the strongest degree of emotional state, as can be seen with the Chinese database.

5. DISCUSSION

In the previous section we investigated the commonality and differences of human perception for perceiving emotion for different languages in the V-A space. In order to discuss the obtained results, we try to answer the following three questions. First, are the neutral positions the same or different among different languages? Second, whether the directions from neutral to other emotional states are similar or not? Third, whether human subjects can estimate the degree of emotional states for different languages i.e. cross-lingually?

Regarding the first question which investigates whether the position of the neutral state is language dependent or not, the results of statistical analysis using ANOVA tests reveal that the neutral positions are significantly different for activation for all subject groups: Japanese, Chinese and Vietnamese. However, the results for valence were statistically different for the Japanese and German database. From Fig. 1, it is clear that the position of neutral is different among the three subjects groups. For example, evaluation for neutral for the Japanese database as shown in Fig. 1(a) shows that Japanese evaluation is in the center of V-A space; however, Vietnamese evaluation is positive as represented by the Green ellipse. On the other hand, Japanese neutral was negatively evaluated by Chinese subjects as represented by the Red ellipse. For the German database the neutral positions for the three subjects were negatively evaluated since German is not the native language for any listeners. Therefore, the answer of the first question is that neutral positions were significantly different among the three subject groups, which indicates that neutral position depends on the subjects' native language.

The direction from neutral state to other emotional states and the horizontal direction towards positive valence were compared for the three subject groups. The results reveal that directions from neutral to other emotional states were similar for the three subject groups for all databases. Thus, the answer of the second question can be summarized as follows: directions from neutral state to emotions on V-A space are similar among subject groups. This result indicates that subjects can recognize emotional categories even without understanding the speaker language.

Moreover, the Euclidean-distance between the neutral state and emotional states, happy, anger and sad, is used as a measure for the degree of emotional state. The results in Tables 2(a), 2(b) and 2(c) indicate that subjects' estimation for the distance between neutral and other emotional states were quite similar when the emotion was expressed by professional such as with the Japanese and German databases. For the Japanese database all emotional styles were very well preformed and responses from the listeners are similar for sad and anger. However, for happy, the responses of the Japanese listener group are weaker than others. The German database is similar to the Japanese database in that the emotional styles were well-performed. Thus, it seem that there are no differences in each emotion style. However, for Sad, the response of the Chinese listener group is stronger than others.

Emotional styles in the Chinese database are not so clear and there are some differences among listeners. Therefore, this database is suitable to discuss listener responses in real-life. For the Chinese database the responses of Japanese were weaker than the Vietnamese responses, which was moderate; however the Chinese response was the strongest. This study suggests that cultural differences are the main reason behind these results because Chinese listeners' responses are stronger and Japanese listener responses weaker, even in the Japanese and German databases. The reason why the Chinese responses are stronger should be investigated in more details in the future. Thus, the answer of the third question is that the degree of emotional state depend on the clearness of the emotional database and the needs of more investigation.

The most important results from this study are that human perception for different languages is identical in the V-A space i.e. the directions from neutral voice to other emotional states are common among languages. However, the neutral positions are different among languages. Thus, directions could be adopted as features for recognizing emotional states in multi-languages. Moreover, it is also important to normalize the features of emotional states by the features of neutral state for each language individually. These findings can be used for adapting emotion recognition system for different languages, which can be used in the recognition part of S2ST system.

6. CONCLUSION

In this paper, we attempted to find the commonalities and differences in human emotion perception for multi-languages on the valence-activation space. Therefore, three emotional speech databases in three different languages, Japanese, German, and Chinese, were analyzed in valence-activation space. Four emotional states, happy, angry, neutral, and sad, were selected from the three databases. Thirty subjects from three different countries, Japan, China and Vietnam, evaluated the three databases in terms of valence and activation. The central positions of all emotional states were calculated by the average of valence and activation. The results reveal that the directions from neutral to other emotions are similar among subjects groups.

It also found that subjects estimation for the degree of emotional state depend on the expressed emotional styles. The perceived degree for all listeners is similar when the emotional styles are clearly expressed by professionals, while the perceived degree varies when the emotional styles are not clear i.e. expressed by non-professional, such as with the Chinese database. The positions of neutral states were significantly different for the three listener groups, which indicates that neutral position depend on the subjects' native language. The most important results are that, directions from neutral state to emotions could be adopted as features to recognize emotional states in multi-languages; moreover, the results of neutral positions indicate that the acoustic features extracted for emotional speech in multi-languages should be normalized by the acoustic features of neutral. This adaption can help us construct a general speech emotion recognition system, which can be used to detect the emotional states of the source language in a speech-to-speech translation system.

7. 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).

8. 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] Szekely, E., Steiner, I., Ahmed, Z. and Carson-Berndsen, J., "Facial expression-based affective speech translation", In: Journal on Multimodal User Interfaces, in press, DOI: 10. 1007/s12193-013-0128-x, 2013.

[4] Elbarougy, R., and Akagi, M., "Cross-lingual Speech Emotion Recognition System Based on a Three-Layer Model for Human Perception", *Proc. Int. Conf. APSIPA ASC*, 2013.

[5] Elbarougy, R., and Akagi, M., "Toward relaying emotional state for Speech-to-Speech translator: Estimation of emotional state for synthesizing speech with emotion", *Proc. Int. Conf. ICSV21*, Beijing, China on 13-17 July 2014.

[6] 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.

[7] 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).

[8] 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.

[9] Russell, J.A., and Geraldine, P. "A Description of the Affective Quality Attributed to Environments", *Journal of Personality and Social Psychology*, **38**(2), pp. 311–322, 1980.

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

[11] Mori, H., Satake, T., Nakamura, M., and Kasuya, H., "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics," *Speech Communication*, **53**, 36-50 2011.