Emotional Speech Recognition and Synthesis in Multiple Languages toward Affective Speech-to-Speech Translation System

Masato AKAGI, Xiao HAN, Reda ELBAROUGY, Yasuhiro HAMADA
School of Information Science
Japan Advanced Institute of Science and Technology
Ishikawa, Japan, akagi@jaist.ac.jp

Junfeng LI
Institute of Acoustics
Chinese Academy of Science
Beijing, China

***Abstract*—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 cnventional approach to S2ST has focused on processing linguistic information only by directly translating the spoken utterance from the source language to the target language without taking into account para-linguistic and non-linguistic information such as the emotional states at play in the source language. In this work, we explore how to deal with para- and non-linguistic information among multiple languages, with a particular focus on speakers' emotional states, in S2ST scenarios called "affective S2ST." In our efforts to construct an effective system, we discuss (1) how to describe emotions in speech and how to model the perception/production of emotions and (2) the commonality and differences among multiple languages in the proposed model. We then use these discussions as context for (3) an examination of our "affective S2ST" system in operation.**

***Keywords*-Speech-to-speech translation (S2ST) system; paralinguistic and non-linguistic information; emotion recognition/synthesis; multiple languages**

I. INTRODUCTION

Wide spread use of the Internet has ushered in a new era of communication. For example, these days communication can be carried out instantaneously regardless of the distance between two parties, even if the other party is on the other side of the world. However, although spoken language is the most direct means of communication among human beings, it is not yet possible to communicate with others directly across the Internet if a common language is not shared. This makes it challenging to construct universal speech communication environments on the Internet.

One approach to this challenge is constructing a speech-to-speech translation (S2ST) system. S2ST is the process by which a spoken utterance in one language is used to produce a spoken output in another language. Conventionally, automatic S2ST consists of three component technologies whereby the spoken utterance is 1) converted into text using an automatic speech recognition (ASR) system, 2) the recognized speech is translated using a machine translation (MT) system into the target language text, and 3) the target language text is resynthesized using a text-to-speech (TTS) synthesizer [1][2].

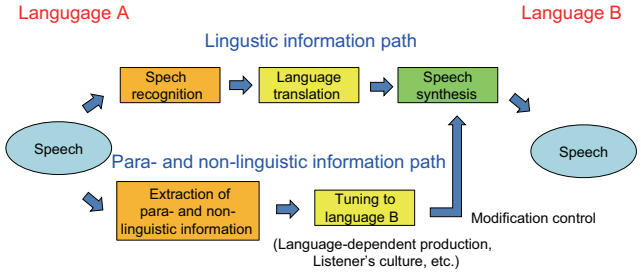

Figure 1. Schematic graph of proposed affective S2ST. This graph contains two paths: one for linguistic information and one for para- and non-linguistic Information.

Speech contains a variety of information [3] including *linguistic*-, *paralinguistic*- and *nonlinguistic*- information. However, conventional S2ST focuses on processing linguistic information only, directly translating the spoken utterance from the source language to the target language, and does not take into account para-linguistic and non-linguistic information such as the emotional states at play in the source language. For example, conventional S2ST systems typically output speech in a neutral voice that remains unchanged even if the input speech changes from one emotional state to another. For natural communication, it is crucial to preserve the emotional states expressed in the source language [4].

In this work, we explore how to deal with para- and non-linguistic information among multiple languages, with a particular focus on speakers' emotional states, called "affective S2ST. " To produce an output of the affective S2ST system colored with the emotional states of the speakers in the source language, the system has to first detect the emotional state at the source language and then convert the acoustic features of the neutral speech produced by the TTS system into those of an emotional speech among multiple languages, as well as to recognize, translate, and synthesize linguistic information in the utterances, as shown in Fig. 1.

In our efforts to construct an effective system, we discuss (1) how to describe emotions in speech and how to model the perception/production of emotions and (2) the commonality and differences among multiple languages in the proposed model. We then use these discussions as context for (3) an examination of our "affective S2ST" system in operation.

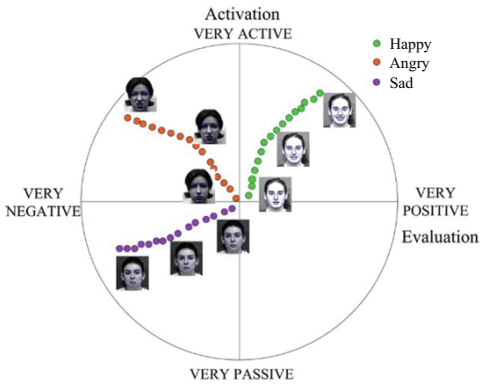

Figure 2. Emotion space spanned by Valence-Activation axes

II. Description of Emotion

A. Emotion Space

Most of the existing techniques for automatic speech emotion recognition/synthesis focus only on the classification of emotional states into discrete categories such as happy, sad, and angry [5][6]. However, a single label or a small number of discrete categories may not accurately reflect the complexity of the emotional states conveyed in everyday interaction [7]. Hence, a number of researchers advocate the use of a dimensional description of human emotion, where emotional states are not classified into an emotion category but rather are estimated on a continuous-valued scale in a multi-dimensional space (e.g., [8][9]).

In this work, we adopt a dimensional description of human emotion, specifically, emotion space spanned by Valence-Activation (V-A) axes [10]. Using the dimensional approach, emotion is represented by a point in this space and emotion categories are represented by regions in an n-dimensional space, where the neutral category lies near the origin, and other emotions lie in a specific region in the n-dimensional space. For example, in the two-dimensional V-A space, happy is represented by a region that lies in the first quarter, in which valence is positive and activation is high, as shown in Fig. 2. Numerical representation is more appropriate to reflect the gradual changes of expressive speech conveyed in everyday interaction, representing the degree within a certain emotion: happy 'not-', 'weak-', 'medium-', and 'strong-happy'. [11]

B. Modeling of Emotion Perception

Scherer [12], in his study of human perception, adopted a version of Brunswik's lens model that was originally proposed in 1956 [13]. In this model, human perception is considered a multi-layer process. In 2008, Huang and Akagi adopted a three-layer model for human perception [14], in which they assumed that human perception for emotional speech is not directly realized from a change of acoustic features but rather from a composite of different types of smaller perceptions expressed by semantic primitives or adjectives describing the emotional voice.

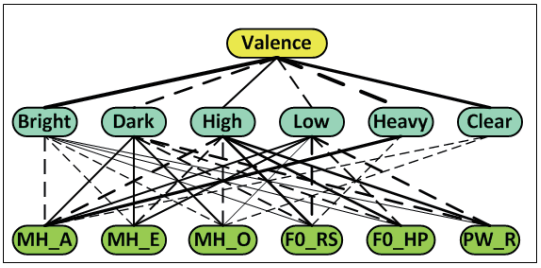

Figure 3. Three layer model for emotion space spanned by Valence-Activation axes. In this figure, Valence is shown [11].

In this work, we adopt a two-dimensional model in [15] based on [14] to represent emotional states using emotion dimensions. Our model consists of three layers, with emotion dimensions as the top layer, semantic primitives as the middle layer, and acoustic features as the bottom layer (Fig. 3).

We took this approach because emotions are not generated in a prototypical modality but rather in complex states that feature a mixture of emotions with varying degrees of intensity. This approach allows a more flexible interpretation of emotions [16].

III. Commonality in Perception of Emotional Speech

Even without an understanding of a certain language, we can still judge the expressive content, i.e., the emotions, of a human voice. To enable communication independent of language, biological features common to both speech production and perception are required [17]. In this section, we discuss commonalities and differences in the perception of emotional speech in the V-A space derived from the results of a listening experiment.

In the experiment, we evaluated the values of valence and activation for three emotional speech databases using three different languages: Japanese, German and Chinese. All three databases were consisted of acted emotions. For the Japanese database, we used the Fujitsu database produced and recorded by Fujitsu Laboratory and selected 20 neutral, 40 happy, 40 angry and 40 sad utterances for a total of 140 utterances. For the German database, we used the Berlin database and selected 200 utterances, 50 of which came from the same four emotional states as the Japanese database. The Chinese database produced and recorded by CASIA using four professional actors (two male and two female) with 48 neutral, 48 happy, 48 angry, and 48 sad utterances selected for a total of 192 utterances.

The three databases were evaluated in terms of valence and activation by 30 subjects: 10 Japanese, 10 Chinese, and 10 Vietnamese. A 5-point scale (-2, -1, 0, 1, 2) was used for the valence and activation evaluations. The valence scale ran from very negative (-2) to very positive (2) and the activation

scale ran from very calm (-2) to very excited (2). The central positions for these emotions were separately calculated by the average value of valence and activation for each emotion category. The central positions of all emotional states were then individually compared for the three listener groups for each database. The results scattered on the V-A space are shown in Fig. 4.

In our analysis of the results obtained on the commonality and differences of human perception for perceiving emotion for different languages in the V-A space, we addressed the following questions: (1) Are the neutral positions the same or different among different languages? (2) Are the directions from neutral to other emotional states similar? (3) Could the subjects estimate the degree of emotional state for different languages, i.e., perform cross-lingual estimations?

For the first question, we found that, according to ANOVA analyses, neutral positions in the V-A space are different among the three subject groups, which indicates that the neutral position is dependent on the subject's native language. For the second question, we found that the directions from neutral to other emotional states were quite similar for the three subject groups of all databases. However, for the third question, no clear tendencies were evident, although the degrees of responses of the Chinese subject group seem larger. More investigation is required to clarify this.

The most significant result here is 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. This demonstrates that direction could be adopted as a feature for recognizing emotional states in multi-languages scenarios. Moreover, it is also important to normalize the features of emotional states by the features of the neutral state for each language individually. These findings can be used for adapting emotion recognition system to different languages.

IV. Multi-lingual emotion recognition/synthesis system

A. Recognition of Emotional Speech: Estimation of position in V-A space

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.

We used an adaptive-network-based fuzzy inference system (AN-FIS) to construct a three-layer model that connects the elements of our recognition system. Each FIS has multiple inputs and one output. Once we obtained the acoustic features set, we constructed an individual estimator to predict the values (-2 to 2; rated by the listening test) of each emotion dimension. For example, in order to estimate the valence dimension using the perceptual model in Fig. 3, we used a bottom-up method to estimate the values (1 to 5; rated by the listening test) of the six semantic primitives

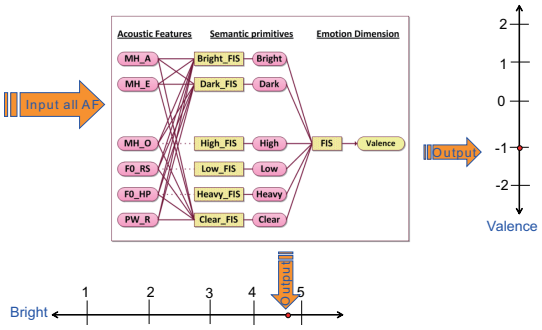

Figure 5. Block diagram of proposed approach for estimating valence based on 3-layer model.

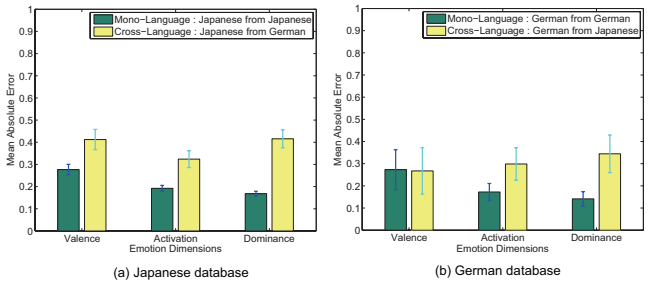

Figure 6. Mean absolute error (MAE) between human evaluation and estimated values of emotion dimensions, in the cases of mono-language and cross-language.

in the middle layer from the six acoustic features in the bottom layer, as shown in Fig. 5. Six FISs were required for this task; one for estimating each semantic primitive. One additional FIS was needed to estimate the value of the Valence dimension from the six semantic primitives. In the same way, the Activation can be estimated using FIS for each semantic primitive. To avoid speaker and language dependency on the acoustic features, we adopt a new acoustic feature normalization, in which all acoustic feature values are normalized by those of the neutral speech.

The mean absolute error (MAE) between the predicted values of emotion dimensions and the corresponding average value obtained from listening tests by human subjects is used as a metric of the discrimination associated with each case. The MAE is calculated by

$$MAE^{(j)} = \frac{\sum_{n=1}^{N} \left| \widehat{x}_i^{(j)} - x_i^{(j)} \right|}{N} \qquad (1)$$

where $j \in \{\text{Valence}, \text{Activation}\}$, $\widehat{x}_i^{(j)}$ is the output of the emotion recognition system, and $x_i^{(j)}$, $-2 \leq x_i^{(j)} \leq 2$ are the values evaluated by the human subjects.

The results, shown in Fig. 6, indicate that the MAEs have small values. Even in the cross-language case, although the mean absolute error of emotion dimensions increased, these increments do not constitute a large difference comparing $x_i^{(j)}$, $-2 \leq x_i^{(j)} \leq 2$.

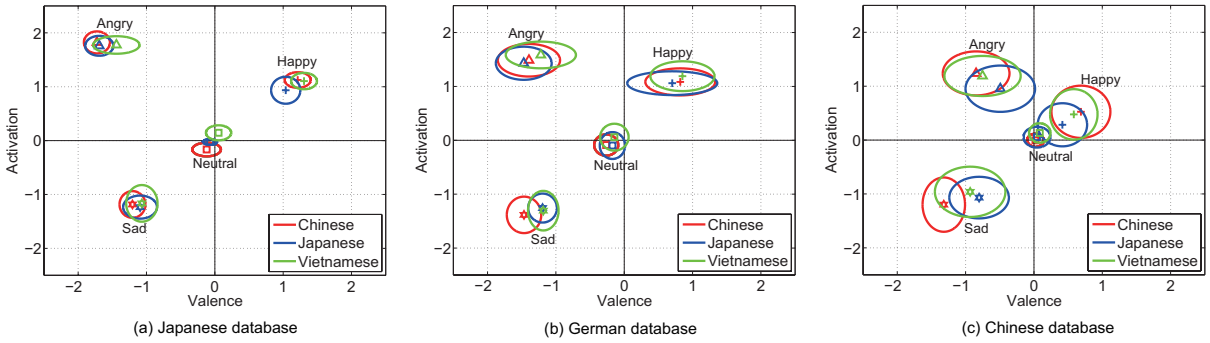

Figure 4. Position of emotional states in valence-activation space.

B. Synthesis of Emotional Speech: Modification of Acoustic Features According to Positions in Emotion Space

Synthesizing emotional speech is an opposite task to recognizing emotion speech i.e., converting a position on an emotion space to the amount of change of the corresponding acoustic features from neutral speech by applying extracted rules from the FISs of the three-layer model used in the emotion recognition (Fig. 5). Although each FIS follows a non-linear mapping, in this paper we linearize the FISs for the rules. STRAIGHT [18] is adopted to synthesize speech using the converted acoustic features. According to internal listening tests, the synthesized speech based on the positions in the V-A space is controlled well and can give the intended impression.

V. Conclusion

In this paper, we discussed how to deal with para- and non-linguistic information among multiple languages, with a particular focus on speakers' emotional states, in S2ST scenarios called "affective S2ST." The system was formulated in the V-A emotional space based on an discussion of commonality and differences of emotion perception among multiple languages. An example of our "affective S2ST" system in operation was also shown.

Acknowledgment

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

References

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

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

[3] H. Fujisaki, "Information, Prosody, and Modeling – with Emphasis on Tonal Features of Speech –," Speech Prosody 2004, 23–26, 2004.

[4] E. Szekely, I. Steiner, Z. Ahmed and J. Carson-Berndsen, "Facial Expression-based Affective Speech Translation," Journal on Multimodal User Interfaces, DOI: 10. 1007/s12193-013-0128-x, 2013.

[5] O. Pierre-Yves, "The Production and Recognition of Emotions in Speech: Features and Algorithms," International Journal of Human-Computer Studies, 59, 157–183, 2003.

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

[7] I. Albrecht, M. Schroder, J. Haber, and H.-P. Seidel, "Mixed Feelings: Expression of Non-basic Emotions in a Muscle-based Talking Head," Virtual Reality, 8, 4, 201–212, 2005.

[8] D. Wu, T.D. Parsons, and S. Narayanan, "Acoustic Feature Analysis in Speech Emotion Primitives Estimation," Proc. InterSpeech 2010, 785–788, 2010.

[9] M. Grimm and K. Kroschel, "Emotion Estimation in Speech Using a 3D Emotion Space Concept," Robust Speech Recognition and Understanding, M. Grimm and K. Kroschel (Eds.), I-Tech, Vienna, Austria, 2007.

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

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

[12] K. R. Scherer, "Personality Inference from Voice Quality: The Loud Voice of Extroversion," European Journal of Social Psychology, 8, 467–487, 1978.

[13] E. Brunswik, "Historical and thematic relations of psychology to other sciences," Scientific Monthly, 83, 151–161, 1956.

[14] C-F. Huang and M. Akagi, "A Three-layered Model for Expressive Speech Perception," Speech Communication, 50, 10, 810–828, 2008.

[15] R. Elbarougy and M. Akagi, "Improving Speech Emotion Dimensions Estimation Using a Three-Layer Model for Human Perception," Acoustical Science and Technology, 35, 2, 86–98, 2014.

[16] Q. Zhang, S. Jeong, and M. Lee, "Autonomous Emotion Development using Incremental Modified Adaptive Neuro-fuzzy Inference System," Neurocomputing, 86, 33–44, 2012.

[17] M. Akagi, "Analysis of Production and Perception Characteristics of Non-linguistic Information in Speech and its Application to Inter-language Communications," Proc. APSIPA2009, 513–519, 2009.

[18] H. Kawahara et al., "Restructuring Speech Representations Using a Pitch Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds," Speech Communication, 27, 187–207, 1999.