Toward Affective Speech-to-Speech Translation: Strategy for Emotional Speech Recognition and Synthesis in Multiple Languages

Masato AKAGI*, Xiao HAN*, Reda ELBAROUGY*†, Yasuhiro HAMADA*, and Junfeng LI‡
*Acoustic Information Science Laboratory, School of Information Science
Japan Advanced Institute of Science and Technology, Japan
E-mail: akagi@jaist.ac.jp
†Department of Mathematics, Faculty of Science, Damietta University, New Damietta, Egypt
‡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 conventional 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 paralinguistic and non-linguistic information such as the emotional states at play in the source language. This paper introduces activities of JAIST AIS lab[1] that explore how to deal with para- and non-linguistic information among multiple languages, with a particular focus on speakers' emotional states, in S2ST applications 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.**

I. Introduction

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 if a common language is not shared. This makes it challenging to construct universal speech communication environments. 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, shown in Fig. 1, automatic S2ST consists of three component technologies whereby 1) the spoken utterance is 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].

Speech contains a variety of information [3] including;

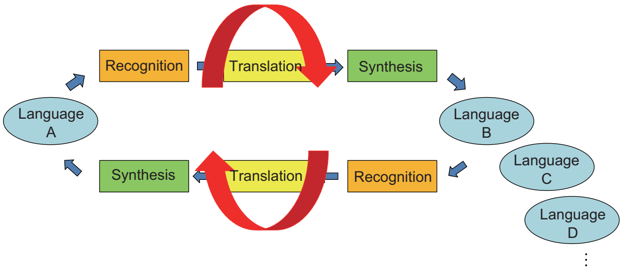


Fig. 1. Schematic graph of speech-to-speech translation (S2ST) system.

- **Linguistic information**: discrete categorical information explicitly represented by the written language or uniquely inferred from context;
- **Paralinguistic information**: discrete and continuous information added by the speaker to modify or supplement the linguistic information; and
- **Nonlinguistic information**: information not generally controlled by the speaker, such as the speaker's emotion, gender, age, etc.

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

[1] Acoustic Information Science Laboratory, School of Information Science, Japan Advanced Institute of Science and Technology

978-616-361-823-8 © 2014 APSIPA

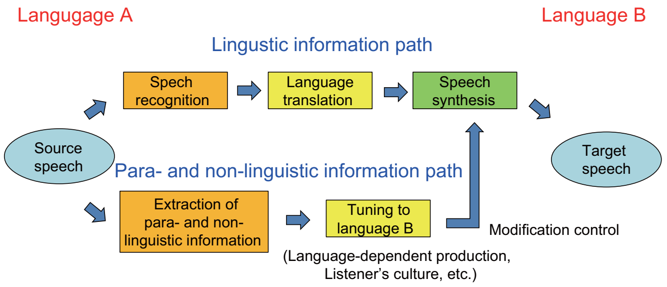

Langugage A
Language B
Lingustic information path
Spech recognition
Language translation
Speech synthesis
Source speech
Target speech
Para- and non-linguistic information path
Extraction of para- and non-linguistic information
Tuning to language B
Modification control
(Language-dependent production, Listener's culture, etc.)

Fig. 2. Schematic graph of proposed affective S2ST. This graph contains two paths: one for linguistic information and one for para- and non-linguistic Information.

recognize, translate, and synthesize linguistic information in the utterances, as shown in Fig. 2.

In our efforts to construct an effective system for "affective S2ST", 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 emotional speech recognition/synthesis system (See Fig. 3) in operation.

II. Description of Emotion

In this section, we introduce emotion space to describe emotions in speech numerically with a two-dimensional space spanned by Valence-Activation (V-A) axes based on dimensional description of human emotion [5], in which emotions are not represented categorically but formulated degree-controllably. Additionally, a three-layer model [6] based on Brunswik's lens model [7] is presented to model perception of emotions from human's perception point of view.

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 [8][9]. 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 [10]. Hence, a number of researchers advocate the use of a dimensional description of human emotion, where emotional states are not classified into an emotion categories but rather are estimated on a continuous-valued scale in a multi-dimensional space (e.g., [11][12][13][14]).

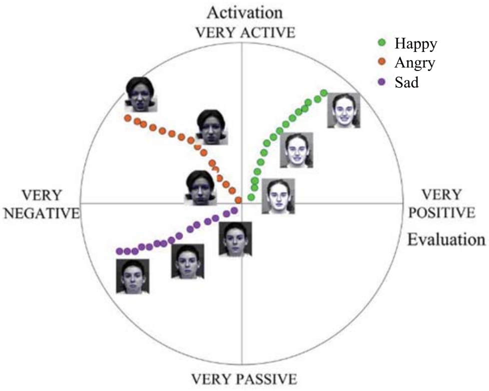


Fig. 4. Emotion space spanned by Valence-Activation axes.

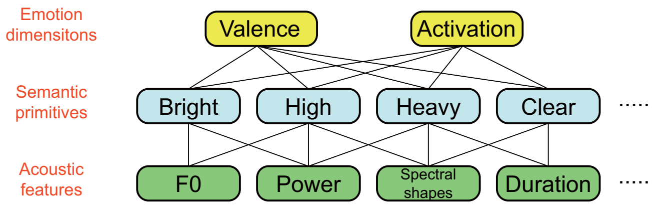


Fig. 5. Three layer model for emotion space spanned by Valence-Activation axes.

In this work, we adopt a dimensional description of human emotion. Using the dimensional approach, emotion is represented by a point in an n-dimensional space and emotion categories are represented by regions in the n-dimensional space, where the neutral category lies near the origin, and other emotions lie in a specific region in the space. For example, in the two-dimensional space spanned by Valence-Activation (V-A) axes [5], 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. 4. Numerical representation is more appropriate to reflect the gradual changes of expressive source and target speech conveyed in everyday interaction, representing the degree within a certain emotion: happy 'not-', 'weak-', 'medium-', and 'strong-happy' [15]. When adopting three-dimensional emotion space, we usually use Valence-Activation-Dominance (V-A-D) axes [16].

B. *Modeling of Emotion Perception*

Scherer [7], in his study of human perception, adopted a version of Brunswik's lens model that was originally proposed in 1956 [17]. In this model, human perception is considered a multi-layer process. In 2008, Huang and Akagi adopted a three-layer model for human perception [18], in which they assumed that human perception for emotional speech is vague and 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.

In this work, we adopt a multi-layer model in [6] based on [18] 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 (Figs. 3 and 5).

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 [19].

.

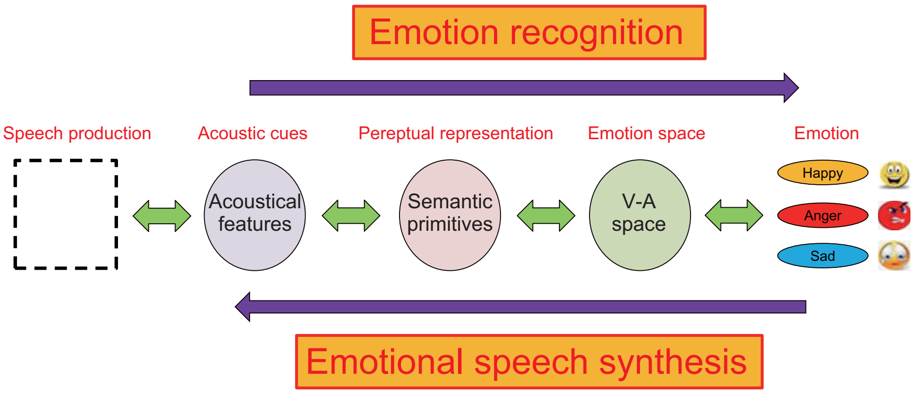


Fig. 3. System for emotional speech recognition/synthesis.

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 [20], that is,

1) Common organ movements for production,
2) Common features produced by common movements,
3) Common impression caused by presenting common acoustic features, and
4) Common behaviors among communicators.

In this section, we discuss commonalities and differences in the perception of emotional speech, mainly for 3), 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). Scatter plots of the responses for all utterances by each database and each listener group are shown in Fig. 6.

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. 7.

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. For example, Japanese listeners feel that Chinese utterances are excited and Chinese listeners feel Japanese utterances are calm. This 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.

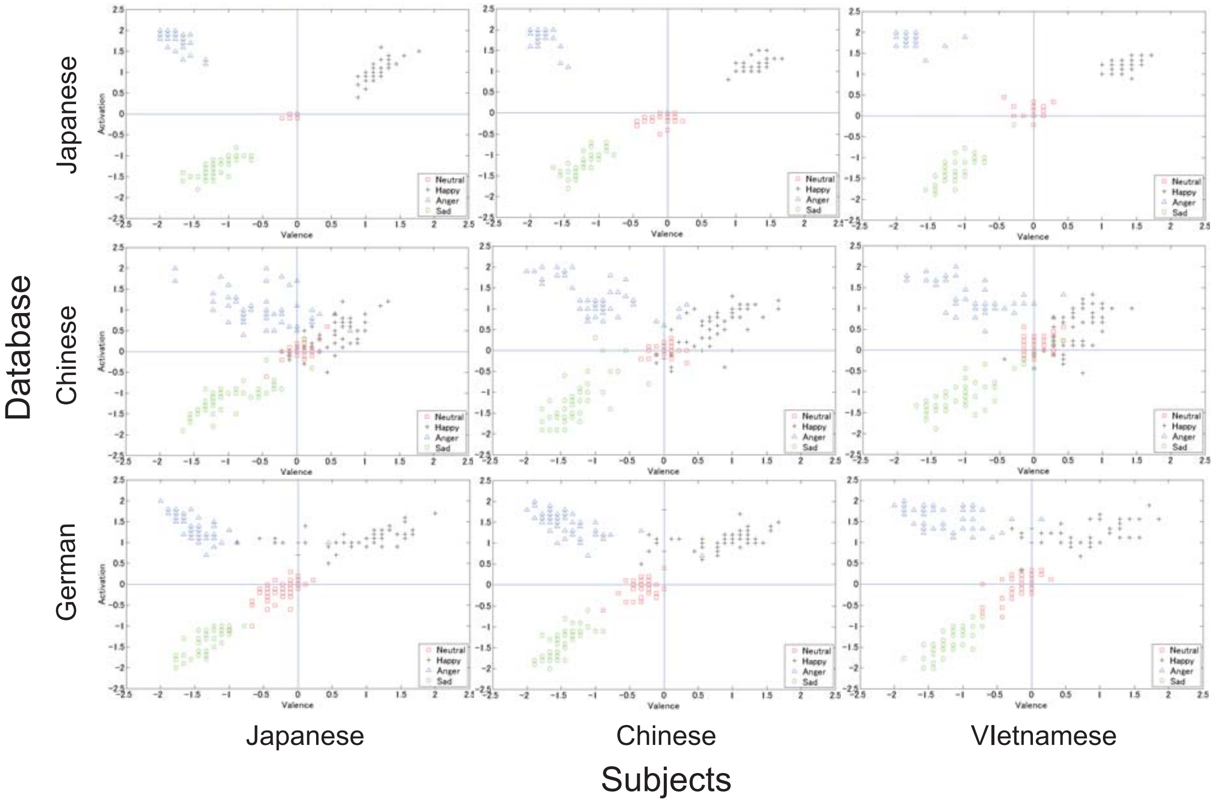
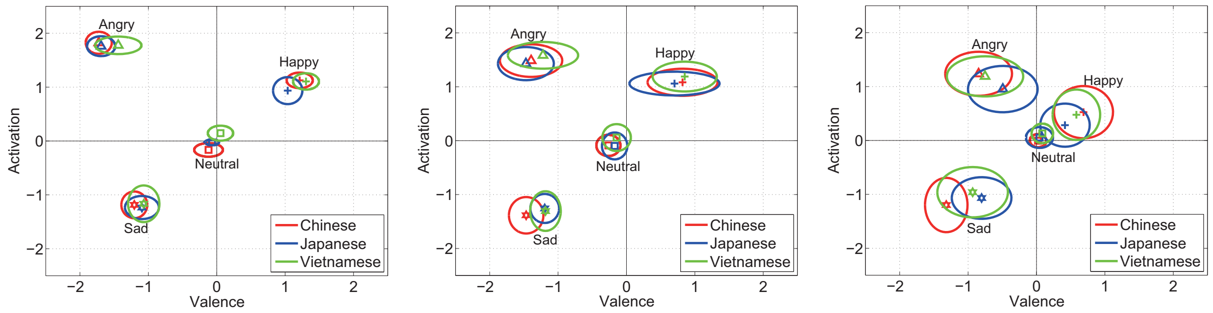


Fig. 6. Scatter plots of responses for all utterances by every database and listener group in valence-activation space.



(a) Japanese database

(b) German database

(c) Chinese database

Fig. 7. Position of emotional states in valence-activation space.

IV. Multi-lingual Emotion Recognition/Synthesis System

Based on the above findings, we introduce examples of the emotion recognition system for estimating positions of uttered speech in the Valence-Activation-Dominance (V-A-D) space [15] and of the emotional speech synthesis system for modifying acoustic features [18], for an affective S2ST system.

A. *Construction of Three-layer Model[15]*

In order to construct a cross-lingual three-layer model to deal with emotion dimensions, valence and activation, we require at least two databases in different languages. The databases and acoustic features used in this work are introduced. Moreover, the semantic primitives and emotion dimensions are evaluated by conducting two listening tests using human subjects as described in next subsections.

1) Speech Materials

In this work, two emotional speech databases were used to construct the model, one in the Japanese language and the other in the German language. The Japanese database is the multi-emotion single speaker Fujitsu database produced and recorded by Fujitsu Laboratory. A professional actress was asked to produce utterances using five emotional speech categories, i.e., neutral expression, joy, cold anger, sadness, and hot anger. The German database is Berlin database [21]. It comprises seven emotional states: anger, boredom, disgust, anxiety, happiness, sadness, and neutral speech. Ten professional German actors (five females and five males) spoke ten sentences with an emotionally neutral content in the seven different emotions.

2) Acoustic Features

We extracted a set of 21 acoustic features which can be grouped in several subgroups:

F0-related features: f0 mean value of the rising slope (F0 RS), the highest F0 (F0 HP), the average F0 (F0 AP) and the rising slope of the first accentual phrase (F0 RS1).

Power envelope-related features: mean value of the power range in the accentual phrase (PW RAP), the power range (PW R), the rising slope of the first accentual phrase (PW RS1), the ratio between the average power in the high-frequency portion (over 3 kHz) and the average power (PW RHT).

Power spectrum-related features: the first formant frequency (SP F1), the second formant frequency (SP F2), the third formant frequency (SP F3), spectral tilt (SP TL), and spectral balance (SP SB).

Duration related features: total length (DU TL), consonant length (DU CL), ratio between consonant length and vowel length (DU RCV).

Voice quality related features: the mean value of the difference between the first harmonic and the second harmonic H1-H2 for vowels /a/, /e/, /i/, /o/ and /u/ per utterance MH A, MH E, MH I, MH O, and MH U.

All the 21 acoustic features were extracted for both the Fujitsu and Berlin databases.

3) Semantic Primitives Evaluation

In this work, we assume that human perception is a three-layer process. It is 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 in [18]. These smaller percepts or adjectives are finally used for detecting the emotional state of the speaker. These adjectives can be subjectively evaluated by human subjects. Therefore, a set of adjectives describing the emotional speech were selected as candidates for semantic primitives. These adjectives are: Bright, Dark, High, Low, Strong, Weak, Calm, Unstable, Well-modulated, Monotonous, Heavy, Clear, Noisy, Quiet, Sharp, Fast, and Slow.

For the evaluation of semantic primitives, we used listening tests. In these tests, the stimuli were presented randomly to each subject through binaural headphones at a comfortable sound pressure level in a soundproof room. Subjects were asked to give scores to each of the 17 semantic primitives on a 5-point scale (‘1–Does not feel at all’, ‘2–Seldom feels’, ‘3–Feels a little’, ‘4–feels’, ‘5–Feels very much’). The 17 semantic primitives were evaluated for the two databases. The scores given by the individual subject were averaged for each semantic primitive per utterance.

4) Emotion Dimensions Evaluation

The two databases were evaluated through the listening tests along the two dimensions of valence and activation. For the emotion dimensions evaluation, a 5-point scale $\{-2, -1, 0, 1, 2\}$ was used: valence (from -2 very negative to +2 very positive) and activation (from -2 very calm to +2 very excited).

5) Selection of Semantic Primitives and Acoustic Features

To accomplish this task, a top-down method was used as follows:

- The correlation coefficients between evaluated values for each emotion dimension (top-layer) and evaluated values of each semantic primitive (middle layer) were calculated;
- The highly correlated semantic primitives were selected for each emotion dimension as an adjective that describes this dimension;
- the correlation coefficients between evaluated values for each selected semantic primitive (middle layer) in the second step and extracted values for each acoustic feature (bottom layer) were calculated; and
- The highly correlated acoustic features were selected for each semantic primitive.

For each emotion dimension, the selected acoustic features are considered to be the features most relevant to the used dimension in the top layer.

6) Selection Results

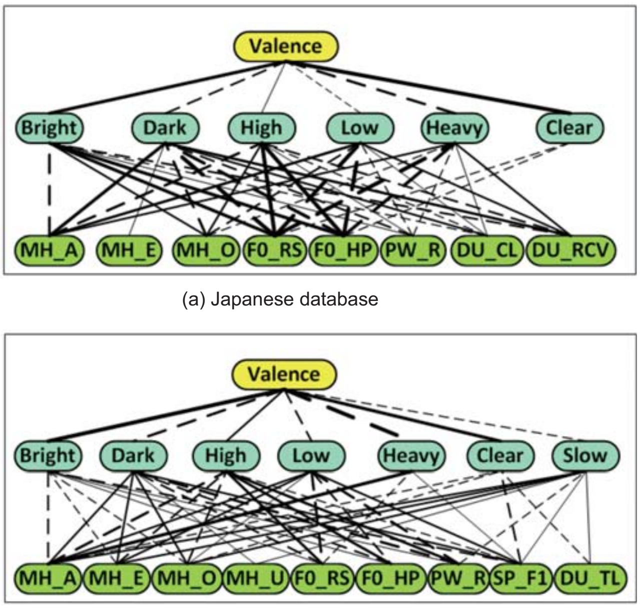
Using the method mentioned above, firstly, the most relevant semantic primitives were selected for each emotion dimension. Secondly, the most relevant acoustic features for each semantic primitive were selected. Finally, a perceptual three-layer model was constructed for each emotion dimension. Figure 8 illustrates the valence perceptual model, where the solid lines in this figure represent a positive correlation, and the dashed ones indicate a negative correlation. The thickness of each line indicates the strength of the correlation; the thicker the line is, the higher the correlation.

In order to construct a perceptual three-layer model for each emotion dimension in a bilingual case, we combined two perceptual three-layer models for the two databases in every dimension individually. The common acoustic features between the two languages were selected to constitute the bottom layer for the bilingual perceptual models. Moreover, the common semantic primitives between the two languages were selected as semantic primitives for the bilingual case. For example, the valence perceptual model for the bilingual case is shown in Fig. 9.

B. Recognition of Emotional Speech: Estimation of position in V-A-D space[15]

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 Figs. 8 and 9, we used



(a) Japanese database

(b) German database

Fig. 8. Valence perceptual model. (a) Japanese database. (b) German database. The solid lines indicates positive correlation, and the dotted ones, a negative correlation. Width of the lines indicates degree of correlation. Simply put, the thicker the line is, the higher the correlation.

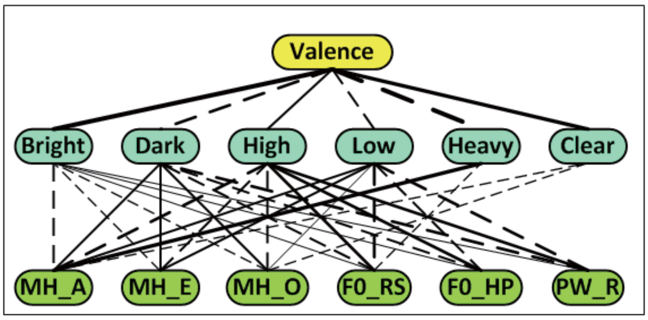


Fig. 9. Bilingual valence perceptual model. The definition of the lines is the same as in Fig. 8.

a bottom-up method to estimate the values (1 to 5; rated by the listening test) of the semantic primitives in the middle layer from the acoustic features in the bottom layer, as shown in Fig. 10. The same number of FISs as the number of semantic primitives was required for this task, i.e., one for estimating each semantic primitive. One additional FIS was needed to estimate the value of the Valence dimension from the 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 an 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.

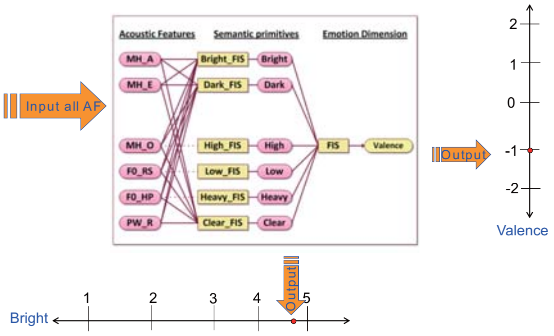
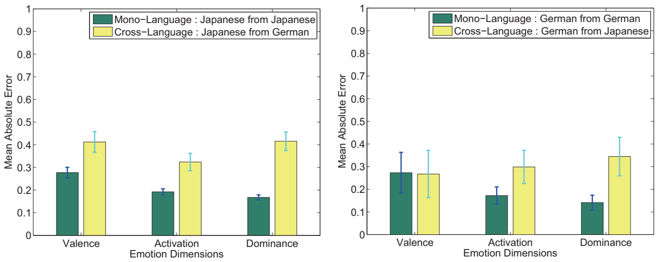


Fig. 10. Block diagram of proposed approach for estimating valence based on 3-layer model for Japanese-German bilingual system (See Fig. 9).



(a) Japanese database

(b) German database

Fig. 11. Mean Absolute Error (MAE) between human evaluation and the estimated values of emotion dimensions, in the case of mono-language and cross-language [15]. 'German from Japanese' in the cross-language case indicates that the German database was processed using the trained FISs with the Japanese database.

The MAE is calculated by

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

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

Figure 11 shows the MAEs. In the figure, for example, 'German from Japanese' in the cross-language case indicates that the German database was processed using the trained FISs with the Japanese database. The results 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 $x_i^{(j)}$, $-2 \le x_i^{(j)} \le 2$.

C. Recognition of Emotional Speech: Results for emotion classification in V-A-D space[22]

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 (Fig. 12(a)). This section also investigates whether the acoustic feature realization of specific emotion is language independent.

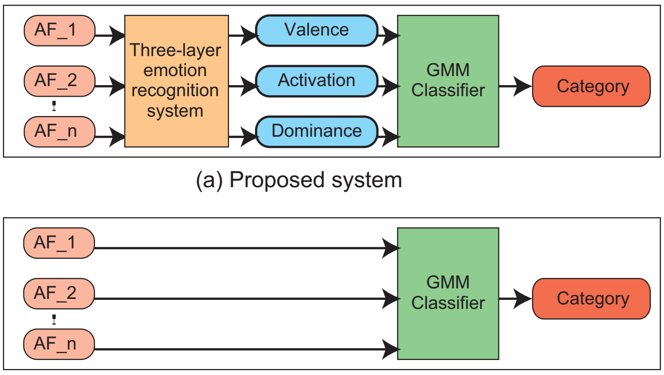


Fig. 12. Emotion classification systems. (a) Proposed system. (b) Traditional system.

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 categories using GMM classier (Fig. 12(b)). Although the GMM classifier is a traditional one, the most important thing in this comparison is how much is the recognition rate using the proposed method improved comparing with that using acoustic features directly.

For investigating the language independent for emotion classification, the performance of the proposed bilingual system was compared with those of the mono-language and cross-language emotion recognition system. In case of mono-language, the system is trained and tested using the same language, and in case of cross-language, the system is trained using one language and tested using the second language. In the bilingual case, the proposed system is constructed by combining two mono-language systems as shown in Section 4-A and the traditional system is trained using two languages. The results of the traditional and the proposed system are shown in Tables I and II for Japanese language and in Tables III and IV for German language.

From Tables I- IV, 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 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 Tables II and IV indicate a small difference between the mono-language, cross-language and the bilingual cases, which reveal that the acoustic feature realization is language independent.

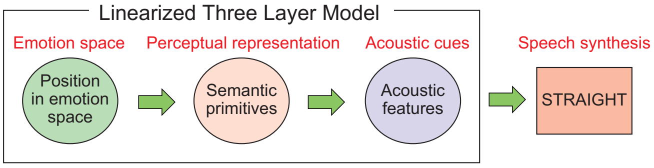


Fig. 13. Schematic graph of emotional speech synthesis.

D. Synthesis of Affective Speech: Modification of Acoustic Features According to Semantic Primitives[18]

Synthesizing emotional speech is an opposite task to recognizing emotion speech (See Fig. 3) i.e., converting a position on the emotion space to the amount of deviations of the corresponding acoustic features of neutral speech by applying extracted rules from the FISs of the three-layer model used in the emotion recognition, as shown in Fig. 13. In this section, we introduce an example of voice conversion from neutral utterances to emotional ones with impressions of 17 semantic primitives [18], to investigate relations between acoustic features and perception of semantic primitives using the synthesized speech. Although each FIS follows a non-linear mapping, in this work, we linearize the FISs to generate conversion rules. STRAIGHT [23] is adopted to synthesize speech using the converted acoustic features.

Why we adopted voice conversion for synthesizing emotional speech instead of directly applying an HMM-base text-to-speech synthesis method learned with categorized emotional speech data, is that we want to fill up the emotion space with synthesized speech continuously.

1) Rule-based Voice Conversion

In our work, we used speech morphing technique to synthesize Japanese affective speech. Our speech morphing process is presented in Fig. 14. Fistly, STRAIGHT[23] is used to extract F0 contour, power envelope, and spectrum of the neutral speech signal while segmentation information was measured manually. Then, acoustic features in terms of F0 contour, power envelope, spectrum, and duration were modified basing on morphing rules inferred from the sets of variation coefficients, mentioned in the next section. Finally, affective speech is synthesized from the modified F0 contour, power envelope, spectrum and duration using STRAIGHT[23]. The conversions are carried out according to the flow presented in Fig. 15.

2) Generation of Conversion Rules

To generate converting rules in the resulting relationship between acoustic features and semantic primitives, we need to obtain the morphing parameters by calculating the difference of acoustic features between the input neutral utterance and the utterances of the intended semantic primitive. Considering the trained FISs to construct a three-layer model, especially focusing on the FISs to estimate the values (1 to 5; rated by the listening test) of the semantic primitives in the middle layer from the acoustic features in the bottom layer, we generated linearized morphing rules. There is one rule for one semantic primitive. One rule has 16 parameters which

TABLE I

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 II

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 III

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 IV

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**

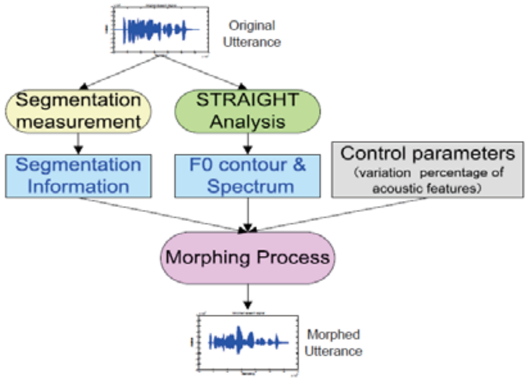


Fig. 14. Process of morphing voices in STRAIGHT.

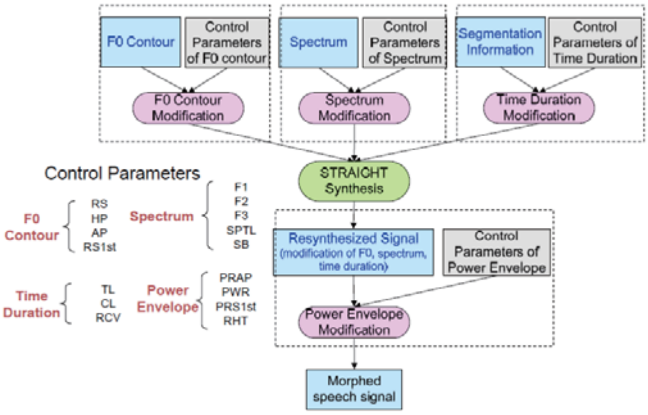

F0 Contour
Control Parameters of F0 contour
Spectrum
Control Parameters of Spectrum
Segmentation Information
Control Parameters of Time Duration
F0 Contour Modification
Spectrum Modification
Time Duration Modification
Control Parameters
STRAIGHT Synthesis
F0 Contour
RS
HP
AP
RS1st
Spectrum
F1
F2
F3
SPTL
SB
Time Duration
TL
CL
RCV
Power Envelope
PRAP
PWR
PRS1st
RHT
Resynthesized Signal (modification of F0, spectrum, time duration)
Control Parameters of Power Envelope
Power Envelope Modification
Morphed speech signal

Fig. 15. Acoustic Feature Modification Process.

control the 16 acoustic features. The values of the parameters are the percentage of changes to an acoustic feature of an input neutral utterance, and the ranges of the values were pre-calculated by the following method. Firstly, we measured the differences between the values of the acoustic features of emotional speech and those of the neutral utterance from which it should be morphed. Then we calculated how much the acoustic features of each utterance varied compared to those of the neutral utterance (i.e., percentage variation) by dividing the differences in the values of the acoustic features with those of the corresponding neutral utterance. Finally, we averaged the percentage variations of each of the utterances in the database to give the values of acoustic features for each semantic primitive.

3) Listening Tests and Results

To examine whether the morphed acoustic features are significant to the perception of semantic primitives or not, an listening test was conducted to evaluate the morphed utterances by comparing them to the neutral utterances from which they were morphed. In this experiment, 68 (17 semantic primitives x 3 degrees) morphed speech utterances were produced by implementing the generated rules for the semantic primitives, giving three morphed speech utterances controlled on degrees of the impressions of each semantic primitives. Neutral speech utterances were in the Japanese Fujitsu database.

In the experiment, Scheffe's method of paired comparison was used to evaluate the intensity of the semantic-primitive. Two stimuli, which were selected two from four (one neutral and three morphed) utterances for each semantic primitives, were presented to the subjects randomly through a binaural headphone at a comfortable sound pressure level. Subjects were ten male Japanese graduate students with normal hearing ability and were asked to evaluate which stimulus (A or B) had a stronger intensity (0 to 2 for B and 0 to 2 for A) of the semantic primitive according to a five-grade scale. The detailed results are shown in Chapter 5 "Verification of the emotional perception model" in [18].

The results of the listening test indicate that (1) most of the morphed speech utterances were perceived as the semantic-primitive intended by the morphed speech utterance, and (2) listeners were able to perceive four levels of intensity for each semantic primitive, except for quiet, for which only three levels were perceived. The difficulty in perceiving different intensity levels for quiet could be because the neutral utterances are intrinsically quiet. These results suggest that the created base rules are effective.

4) Future works

In this section, we introduced a method to give listeners impressions of 17 semantic primitives by modifying the appropriate acoustic features. In the next steps, (1) we will propose a model to estimate which acoustic features are selected and how much the selected acoustic features are modified from those of neutral utterances, according to positions of semantic primitives (from 1 to 5), and (2) we will propose another model to estimate the positions of semantic primitives (from 1 to 5) from desired positions in the V-A or V-A-D emotion space. Combining these two models, we will construct a emotional speech synthesis system according to the position in the emotional space. This is the strategy the authors are following.

Rules for relationship between the middle layer (semantic primitives) and the top layer (positions in emotion space) in the three-layer model are still under investigation and are tuned carefully. 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. In future conferences, we will present better results.

V. Conclusions

In this paper, we introduced 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 or V-A-D 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 a 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] 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] 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.

[6] 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.

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

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

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

[10] 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.

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

[12] M. Schroder, R. Cowie, and E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," Proc. Eurospeech 2001, 87–90, 2001.

[13] 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.

[14] I. Kanluan, M. Grimm, and K. Kroschel, "Audio-Visual Emotion Recognition using an Emotion Space Concept," Proc. EUSIPCO 2008, 2008.

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

[16] C. Breazeal, " Emotion and sociable humanoid robots," International Journal of Human-Computer Studies, 59, 119–155, 2003.

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

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

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

[20] M. Akagi, “Analysis of Production and Perception Characteristics of Non-linguistic Information in Speech and its Application to Interlanguage Communications,” Proc. APSIPA2009, 513–519, 2009.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A Database of German Emotional Speech,” Proc. Interspeech2005, Lissabon, Portugal, 2005.

[22] R. Elbarougy and M. Akagi, “Toward Relaying Emotional State for Speech-to-Speech Translator: Estimation of Emotional State for Synthesizing Speech with Emotion,” Proc. 21st Int. Conf. Sound and Vibration, 2014.

[23] 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.