2015 RISP International Workshop on Nonlinear Circuits,
Communications and Signal Processing (NCSP'15)
Kuala Lumpur, Malaysia, February 27 - March 2, 2015

A study on perception of emotional states in multiple languages on Valence-Activation approach

Xiao Han[1], Reda Elbarougy[2], Masato Akagi[1],Junfeng Li[3], Thi Duyen Ngo[4] and The Duy Bui[4]

[1] JAIST
1-1, Asahidai, Nomi,
Ishikawa, 923-1292 Japan
han_xiao@jaist.ac.jp
akagi@jaist.ac.jp

[2] Damietta Univ.
Damietta, 3451 Egypt
Elbarogy2000@yahoo.com

[3] IOACAS,China
No.21 North 4th Ring Road,
Haidian District, Beijing,
100190 China
lijunfeng@hccl.ioa.ac.cn

[4] VNU-UET, Hanoi
144 Xuan Thuy Street,
Cau Giay,Hanoi, Vietnam
duyennt@vnu.edu.vn
duybt@vnu.edu.vn

Abstract

Human beings can judge emotional states of a voice only by listening, no matter thay understand the language or not. Investigating the commonalities and differences of emotional states perception among multiple languages is important to understand how human beings perceive emotional states among multiple languages, and to build a human perception model independent to used languages. This paper investigates commonalities and differences among multiple languages in human perception of emotional states based on a dimensional approach. The results suggest that human beings can perceive emotional states regardless of languages. Moreover, the results can be used to build a human perception model regardless of languages only by control the deviations between neutral voices and other emotional states.

1. Introduction

Communication is a essential part of human beings' social life, and speech is one of the most common way for us to communicate with others. From the experience in daily life, it is found that human beings can judge the emotional states of a voice only by listening. This kind of situation occur not only when they listen to their native language, but also when they listen to non-native languages they do not familiar with. It suggests that there is another way to communicate each other without common language. The commonalities and differences of emotional states perception among multiple languages are parts of the fundamental knowledge on how human beings perceive emotional states of different languages. Since it can help us to understand how human beings perceive emotional states among multiple languages, and can guide us to build a human perception model independent to used languages, it is significantly important for us to investigate.

The purpose of this study is to investigate commonalities and differences among multiple languages of human perception for emotional states from speech signal. In order to achieve this goal, some detailed informations about the commonalities and difference, such as the neutral voices, the relationship of neutral state to other emotional states and the degree of emotional states, are investigated.

In fact, many previous studies [1, 2, 3, 4] have concerned with comparison of speech emtion perception among different languages. In these studies, two kinds of emotion disctiptions have been used to present emotional states. They are categorical approach and dimensional approach. Both of the two emotion discriptions can present emotional states clearly. Since the dimensional approach can be used to present not only category of emotional states, but also degree of emotional states [5, 6], a 2-dimension approach, Valence-Activation approach is adopted in this study.

2. Experiment

In the study, commonalities and differences among multiple languages for human perception of emotional states in speech are investigated on Valence-Activation approach. To compare the commonalities and differences among multiple languages, a listening test was carried out. In the listening test, thirty listeners from three different countries were asked to evaluated emotional contents in five different languages.

2.1 Databases

Five emotional speech databases consisted of acted emotions in five different languages are selected in the listening test. The five databases are CASIA database, IEMOCAP database, Berlin database [7], Fujitsu database and VNU database in Chinese, Amarican English, German, Japanese and Vietnamese respectively. The utterances we used are covered four basic emotion, neutral, happyness, anger and sadness. The total number of utterances selected are 556.

2.2 Subjects

In the listening test, all utterances are evaluated by 30 listeners from three different countries and speak three different native languages. Ten Chinese, ten japanese and 10 Viet-

namese are asked to take part in the listening test. All of them are graduate students from 20 to 35 years old. No listeners have hearing impairment.

2.3 Procedure

In order to obtain the position of each utterance on Valence-Activation approach, All utterances are evaluated by 3 listener groups in the listening test. Every listener is required to evaluate the emotion in the voices by his/her own perceived impression based on the way of speaking, but not on the content itself. A seven-point-scale are used for their evaluation. For valence axis, the seven points are as very negative (-3), mid negative (-2), low negative (-1), neutral (0), low positive (1), mid positive (2) and very positive (3), and for activation axis, they are very calm (-3), mid calm (-2), low calm (-1), neutral (0), low excited (1), mid excited (2) and very excited (3).

The listening test was divided into 2 sessions and 4 parts in each session. In the two sessions, listeners are asked to evaluate the scores of valence axis and activation axis independently. In each session, four parts, introduction [8], training, pre-test and main-test, are included. The previous three parts are aimed to help listeners to understand the basic theory of dimensional representation and check their understanding and abilities. In the main-test, a MATLAB GUI was used to input given scores for evaluation. All 556 stimuli were presented randomly through a binaural headphones (STAX SPM-a/MK-2) at a comfortable sound pressure level in a soundproof room. The D/A device and driver are M-AUDIO Fast Track Pro and Asio sterio driver.

3. Results

To camparie the experiment results on Valence-Activation approach, every emotional state is presented as ellipse distributed on the space. Coordinate (x_E, y_E) presents the center of the ellipse, in which x_E and y_E are the averages of valence and activation of the emotional states (E). Moreover, the standard deviations of valence and activation are presented by the horizontal and vertical radii of the ellipse. Figures 1 (a)-(e) show the positions of the four emotional states on Valence-Activation approach. For each database, colors red, blue and green represent the evaluated results of three subjects groups, Chinese, Japanese and Vietnamese, respectively.

3.1 Position of neutral voice

In order to discuss the position of neutral voice, we extract the ellipse of neutral states. Figure 2 (a)-(e) show the position of neutral states. For each database, colors red, blue and green represent the evaluated results of three subjects groups, Chinese, Japanese and Vietnamese, respectively. From the figures, we can see that the position of neutral voice are not significantly different among three listener groups for all databases.

Table 1: Angles of the vector from neutral state to other emotional states.

(a) Chinese Database

Subject	J	C	V
Neutral-Happy	46.2°	41.9°	48.9°
Neutral-Angry	136.0°	138.6°	141.6°
Neutral-Sad	222.2°	220.5°	225.2°

(b) English Database

Subject	J	C	V
Neutral-Happy	40.7°	49.4°	59.3°
Neutral-Angry	134.3°	141.1°	155.9°
Neutral-Sad	231.1°	231.9°	234.4°

(c) German Database

Subject	J	C	V
Neutral-Happy	50.2°	48.4°	54.4°
Neutral-Angry	142.5°	142.4°	140.3°
Neutral-Sad	232.8°	225.8°	227.5°

(d) Japanese Database

Subject	J	C	V
Neutral-Happy	40.0°	35.6°	44.3°
Neutral-Angry	136.1°	137.6°	138.1°
Neutral-Sad	228.7°	228.2°	227.4°

(e) Vietnamese Database

Subject	J	C	V
Neutral-Happy	53.5°	46.1°	41.7°
Neutral-Angry	145.9°	149.3°	154.6°
Neutral-Sad	233.8°	229.5°	244.3°

3.2 Direction of emotional states

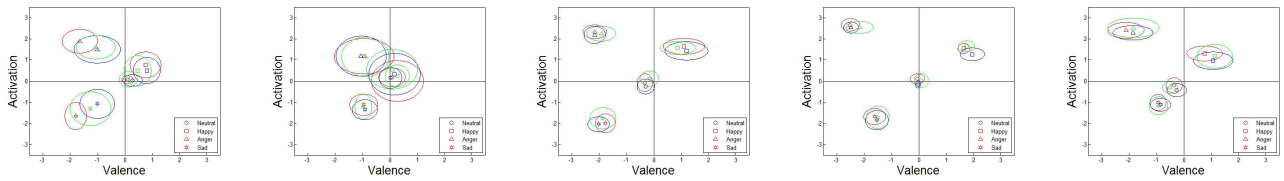
Direction of one emotional state is represented by angle of the vector from the center of neutral state to the center of the emotional state. It is calculated by the following equation:

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

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.

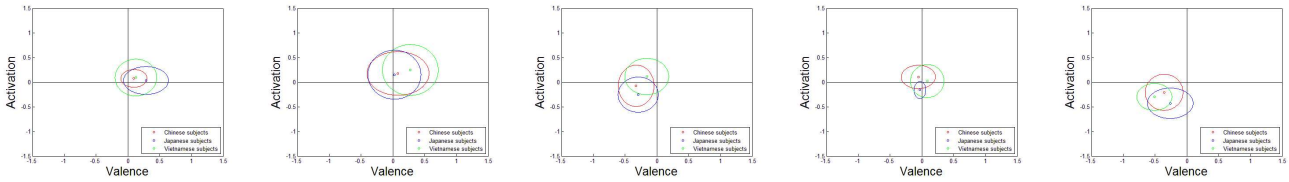
The calculated results are listed in Table 1, and it reveals that directions from neutral state to other emotion states on Valence-Activation approach are similar among three listener groups.

3.3 Distance



(a) Chinese database (b) Enlish database (c) German database (d) Japanese database (e) Vietnames database

Figure 1: Emotional states' position on Valence-Activation approach.



(a) Chinese database (b) Enlish database (c) German database (d) Japanese database (e) Vietnames database

Figure 2: Position of neutral states on Valence-Activation approach.

To calculate degrees of emotional states, distances from neutral to other emotional states on V-A space are used as a metric. Large distance means strong emotional states and vice-versa [**?**]. The Euclidean-distances between neutral to other emotional states on Valence-Activation approach are calculated by the following equation:

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

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.

Table 2 listed the results. It is found that the distance are significantly different among three listener groups.

4. Disccussion

In the previous section, in order to investigate the commonalities and differences among multiple languages on Valence-Activation approach, the results of the listening test was compared in three points of views. They are the position of neutral voice, the direction of emotinal states and the distance of emotional states.

Firstly, for the position of neutral states, the results of listening test are shown in Figure 2 (a)-(e). The figures show that the position of neutral voice are not significantly different among the three listener groups for all databases. Moreover, the position of neutral voice evaluated by the three listener groups are not located at the center of the approach. Considering the reason why the situation occurred in the five databases, there are three hypotheses. The first hypothesis is the position of neutral states in these languages are not at the center of approach. For Chinese, Japanese and Vietnamese databases, since the responces of native speakers are not at the center either. The probablity that the deviation between the position of neutral states and the center of the approach caused by then languages is very low in the three databases. On the other hand, because there are no native speakers of English and German, we cannot know whether the hypothesis influent the results or not. The second hypothesis is that the personality of the speaker influent the results. For Chinese, English and German databases, the number of speakers in one database are more than four, and for Japanese and Vietnamese databases, there are only 1 or 2 speakers. Therefore, the influence of the speakers' personality should be strong in Japanese and Vietnamese databases, but not strong in other three databases. Then, the third hypothesis is that when speakers recording neutral voices, they spoke with other emotional states. This hypothesis would give strong influence to the results in the five databases.

The second point of view is about the direction of emotional states, which is presented by the angle of the vector from the center of the neutral state to the centers of other emotional states. As Table **??** shows, directions of emotinal states on Valence-Activation approach are similar among three listener groups. It supposes human beings can perceive emotional state using these angles no matter they understand the languages or not.

Thirdly, from Figure 1 (a), in Chinese database, the responces of Chinese listeners are stronger than the responces of Japanese and Vietanmese subjects. Similarly with Chinese database, in the Figure 1 (b) and (c), the responces of Japanese listeners are the stronggest in Japanese database, and the responces of Vietnamese listeners are the stronggest in Vietnamese database. Moreover, Table 2 shows the cal-

Table 2: Distances between neutral state and other emotional states.

(a) Chinese Database

Distance	C	J	V
Neutral-Happy	0.96	0.70	0.52
Neutral-Angry	2.52	1.96	1.92
Neutral-Sad	2.57	1.71	1.97

(b) English Database

Distance	C	J	V
Neutral-Happy	0.26	0.23	0.37
Neutral-Angry	1.43	1.72	1.53
Neutral-Sad	1.66	1.53	1.19

(c) German Database

Distance	J	C	V
Neutral-Happy	2.24	2.25	1.79
Neutral-Angry	3.03	3.08	2.78
Neutral-Sad	2.40	2.49	2.54

(d) Japanese Database

Distance	J	C	V
Neutral-Happy	2.24	2.45	2.32
Neutral-Angry	3.55	3.64	3.38
Neutral-Sad	2.37	2.27	2.35

(e) Vietnamese Database

Distance	J	C	V
Neutral-Happy	1.88	1.93	2.21
Neutral-Angry	3.15	3.15	3.12
Neutral-Sad	1.00	0.91	0.96

culated results of the distance between neutral states and other emotional states. For Chinese, Japanese and Vietnamese database, the distances evaluated by native speakers of the language are almost the largest one in the three listener groups. It suggests that native speaker always gives stronger responce than other people. On the other hand, for English and German database, since the three listener groups are not the native speakers of these two languages, it is difficult to distinguish which listener group gave the stronggest response.

5. Conclusions

In this paper, we attempted to investigate the commonalites and differences among multiple languages on Valence-Activation approach for human perception of emotional states in speech. In order to find the commonalies and the differences, we compared the position of neutral state, the direction and the degree of emotional states among the listeners in different native languages on Valence-Activation approach. According to the analysis, we achieved that the commonalities are the position of neutral state and the direction from neutral states to other emotional state, and the difference is the degree of emotional state. Moreover, the results suggest that when human beings perceive emotional states among multiple languages, although the perception of neutral states are different among different listener groups, they can still perceive emotional states regardless of languages. Moreover, the results can be used to build a human perception model regardless of languages only by control the deviation between neutral states and other emotional states.

Acknowledgment

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

References

[1] C. F. Huang, *et al.* "Comparison of Japanese expressive speech perception by Japanese and Taiwanese listeners", *Acoustics2008*, Paris, pp. 2317–2322, 2008.

[2] J. W. Dang, *et al.* "Comparison of Emotion Perception among Different Cultures", *Acoust. Sci. & Tech.*, **31**, 6, 2010.

[3] H. R. Pfizinger, *et al.* "Cross-language Perception of Hebrew and German Authentic Emotional Speech", *ICPhS XVII*, Hong Kong, 17-21, August, 2011.

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

[5] M. Schroder, "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotion", in *ADS 2004*, E. Andre *et al.*, Eds. (Springer, Berlin/Heidelberg, 2004), pp.209-220.

[6] M. Grimm and K. Kroschel, "Emotion Estimation in Speech Using a 3D Emotion Space Concept", in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. (I-Tech Education and Publishing, Vienna, 2007), Chap.16.

[7] F. Burkhardt, *et al.* "A Database of German Emotional Speech", *Proceedings of Interspeech*, Lissabon, Portugal, 2005.

[8] H. Mori, *et al.* "Constructing a spoken dialogue corpus for studying paralinguistic information in expressive conversation and analyzing its statistical/acoustic characteristics", *Speech Communication*, **53**, 36-50, 2011.