Title	Emotion Recognition in Multiple Languages using a Three Layer Model
Author(s)	李，興風
Citation	
Issue Date	2016-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/13638
Rights	
Description	Supervisor:Masato Akagi, School of Information Science, Master

Emotion recognition in multiple languages using a three layer model

Xingfeng LI (1310211)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 10, 2016

Keywords: speech emotion recognition, emotion dimension, human perception, three layered model.

Speech is a complex signal involving information about message, speaker, language, emotion, and etc. It is the fastest and the most natural means of communications between humans in our daily life. This fact has motivated researchers to think of speech as a fast and efficient approach of interaction between human and machine. Most existing researches study on speech recognition, which refers to the process of converting the human speech into a sequence of words. However, closing the gap between human and machine, despite the great process made in speech recognition, it is still a scientific challenge. This is due to among human conversations, non-verbal communication always carries an important information like intention of the speaker. Besides the message conveyed through text, the manner in which words are spoken, conveys essential nonlinguistic information. The same textual message would be conveyed with different meaning by incorporating appropriate emotions. Therefore, it is meaningful to make the machine know human emotions.

Speech emotion recognition (SER) has several applications in our lives. It is useful for enhancing naturalness in human-machine interaction. SER may be used in an on-board car driving system, where information about mental state of the driver may be used to keep him alert during driving. This helps to avoid some accidents, caused by stressed mental state of the driver. It also be useful in web movies and computer tutorial applications where the response of those systems to the user depend on the detected emotions. Beyond these, it is particularly helpful for affective speech to speech translation (S2ST) system in which the spoken utterance in the source language is translated into the target one, and here the translated speech is colored with the same emotional states that conveyed in the original speaker's message. Such kinds of systems are of importance as for promoting cultural exchange activities to foster mutual understanding around the world.

An important issue is the appropriate approach to describe emotion. Most investigations on SER using the speech signal tries to distinguish a small number of emotion categories, like anger, disgust, fear, joy, sadness, surprise, and neutrality. However, generally, expressions of emotion in natural speech may carry any arbitrary value in between, such as little happy, normal happy, or very happy. To describe such degrees or intensities of emotion, it is widely believed that, emotion can be characterized in two dimensions: activation and valence, activation describing the excitation on a scale from calm to excited, and valence describing the negative vs. positive of an emotion. Given that in dimensional approach, emotion can be easily captured as a point, where the numerical description is more suitable to reflect the gradient nature of emotion expression. In our study, we utilized dimensional approach to represent emotion.

From perspective of dimensional approach, there are some studies have been devoted to the analysis of estimating emotion dimensions from acoustic features directly by using different kinds of classifiers. However, owing to the lack of relative acoustic features to valence dimension, the prediction of valence always performed poorly in these studies. To overcome this critical limitation, Elbarougy adopted a three layered model in dimensional method for human perception described by Scherer (Scherer, 1978) and developed by Huang and Akagi (Huang and Akagi, 2008). They assumed that human emotional recognition is not from acoustic features directly, but from a smaller perception by describing emotional voice using adjectives like, bright, dark, fast, or slow, etc. This three layer model consists of acoustic features in the bottom layer, semantic primitives in the middle layer and emotion dimensions in the top layer. Experimental results revealed that such a human perception based three layer model can effectively improved the emotion dimensions estimation in monolingual case.

These days, most of existing studies focus on investigating many speech features and their relationship to emotion. Simultaneously, there are also attempts to employ different feature selection techniques to find the best features for this task. However, the conclusions from different studies are inconsistent. Several studies have shown evidence for certain universal attributes for both speech and music, not only among individuals but also cross cultures. To investigate whether emotional states can be recognized universally or not, Elbarougy proposed a bilingual SER system using dimensional approach derived from the three-layered model. This bilingual perceptual model was constructed with combined information (common acoustic features and common semantic primitives) between two different databases, in Japanese and German. Moreover, normalization between languages was done in the acoustical feature layer to avoid language dependent. Unfortunately, it is found that this model can only work for several language pairs, applications of normalization and common features selection methods into the bilingual emotional speech recognition system resulted degradation of accuracy in emotional states estimation compared with mono-language cases. Due to these limitations of feature selection strategy and languages normalization, emotion dimensions estimation on multiple languages scenario is particularly challenging.

In 2015, commonalities and differences of emotion perception across multiple languages have been investigated by carrying out human listening tests by Han and Elbarougy. Obtained results from that study indicating that direction and distance from neutral to other emotions are similar among languages. Motivated by this new finding, adopting directions and degrees in the emotion dimensional space as distinguished features among emotions, the multiple language SER system can be easily implemented with the assumption that the generalized SER system be able to precisely estimate emotion dimensions in multiple languages. Hereafter, the final goal of this study is to build a multiple language SER system, which be able to accurately predict emotion dimensions such as human response in the dimensional space.

To achieve this system, from perspective of restoring human emotion perceptual processing, we address three important aspects in SER on multiple language scenario: (1) the strategy of powerful feature selection, (2) multinational emotions modelling, (3) classification technique cross cultures.

Using existing emotional speech corpus (Japanese, German, Chinese), we firstly propose a correlation-based features selection procedure, in which correlation coefficients between the emotion dimensions and semantic primitives are calculated in the first step, subsequently, the correlation coefficients between semantic primitives and acoustic features are calculated. Secondly, by mimicking the human perception, all emotion dimensions of three databases are trained and tested based on the three layered model with fuzzy inference system (FIS) simultaneously. Finally, with accurately estimated emotion dimensions, we apply the knowledge of commonalities and differences of human perception among languages to classify emotional categories using direction and degree by SVM.

Eventually, a comparative analysis of the performance of multilingual SER system is studied. Comparisons of multilingual SER, monolingual SER, and bilingual SER are fully discussed. With the investigation on SER using proposed strategies, it indicates that the estimation of emotion dimensions of multilingual SER can achieve a well performance just as human beings do. Even compared with monolingual and bilingual SER, the estimated results are really evenly matched.