Speech Emotion Recognition in Multiple Languages using a Three Layered Model

Xingfeng Li, Masato Akagi
[1, 2]Japan Advanced Institute of Science and Technology

Abstract: ***Recent studies on speech emotion recognition (SER), as performed in most works so far, generally can be simply analyzed by each single language. Human emotion perception is cross cultures and an automatic SER system should be able to recognize it such as. We therefore present a SER system in multilingual scenario from perspective of human perceptual processing. The goal is twofold. Firstly, to predict multilingual emotion dimensions accurately as well as human annotated. To this end, a three layered model consists of acoustic features, semantic primitives, and emotion dimensions, along with Fuzzy Inference System (FIS) are studied. Secondly, by the knowledge of commonalities and differences of human perception among languages in dimensional space, a language normalization approach by extracting direction and degree is adopted to detect multilingual emotions. Results proved that estimation performance of emotion dimensions comparable to human evaluation is furnished, and classification rates that are close to monolingual SER system performed are achieved.***

Keywords: emotion recognition in speech, three layered model, emotion dimension.

1. Introduction

Speech processing is widely studied in the area of affective computing to enable computers possess sufficient intelligence to understand human behavior. Most automatic speech systems focused on the process of natural language understanding by capturing the linguistic contents of spoken utterances. Such language understanding can be further improved if the state of emotion of the speaker can be detected. Automatic SER can be very indispensable for applications which require natural human-machine interaction such as Affective Speech to Speech Translation system in which the expressed emotions are of importance in the communication among subjects with different cultures. Many works on monolingual SER using different single corpus haven been achieved in the past several decades, however, human emotion perception is proved to be cross culture even without understanding of that language used. And an automatic SER system hence is expected to be able to recognize emotions such as humans. This research aims at constructing a SER system that can be analyzed by multiple languages.

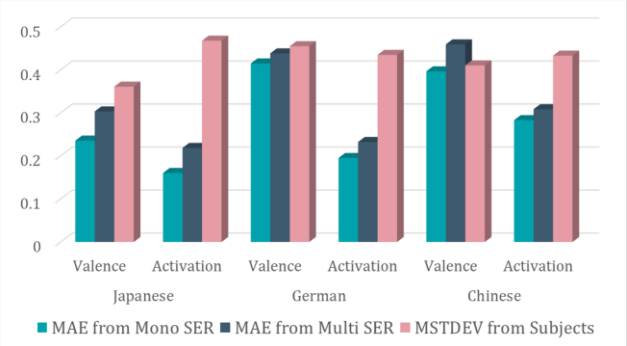

Figure 1. Comparison of performance of estimation of emotion dimensions by System and Human Subjects. MAE is short for mean absolute error, and MSTDEV stands for mean standard deviation among human beings.

Table 1: Classification rates of each single language by Mono-lingual SER (Mono), Multilingual SER (Multi), and Referenced studies after [13] [14] [15] respectively for Fujitsu database, Berlin Emo-DB, and CASIA corpus.

[%]	Mono	Multi	Referenced
	Fujitsu Database		
Neutral	100	95	80
Joy	97.5	93	97.5
Anger	95	95	92.5
Sad	95	100	100
Average	96.88	95.75	92.5
	Berlin Emo-DB		
Neutral	98	90	88.5
Joy	86	82	69
Anger	86	82	93.7
Sad	90	96	94.3
Average	90	87.5	86.38
	CASIA emotional corpus		
Neutral	96	89	98
Joy	97	77	82.25
Anger	88	73	90.25
Sad	92	92	94.5
Average	93.25	82.75	91.25

2 Conclusion

We achieved a multilingual SER from perspective of restoring the processing on human emotional perception by a three layered model. Estimation of emotion dimensions which regardless of cultures is studied. With the knowledge of human perception among languages in V-A space, we adopted direction and degree as features to recognize emotion in multilingual scenario. Experimental results show that proposed multilingual system has the ability to precisely estimate emotion dimensions as human annotated, and classification performance of proposed system close to that achieved by monolingual SER can be furnished.