論 文 の 内 容 の 要 旨

論文題目　Research on Emotion Recognition using Speech and Physiological Signals
（音声・生体情報を用いた感情識別手法に関する研究）

氏　名　張 皓

1. Background

Most developed countries in modern society are facing serious problems with the increasing number of lifestyle related diseases, which are greatly influenced by negative emotions and unstable emotional states. The literature has revealed a strong relationship between emotional experiences and human health. Negative emotions have deleterious effects on health in patients, while positive emotions play a protective role in keeping people from diseases. Besides the direct influence of physical or mental health, they also engage in behaviors that damage one's health such as alcoholism, smoking, drug abuse, etc.

Recently, many engineering methodologies to recognize emotions automatically have been proposed due to the growing needs of healthcare service. As speech is the most natural and efficient type of human communication, researchers are still searching a better way to model human emotion using speech signals. It's assumed that utterances (phrases, short sentences, etc.) are the fundamental units for emotion recognition in current researches. However, it has been questioned because of the difficulties of avoiding the influence from spoken content and style in using the utterance-wise features. Moreover, valuable but neglected information could be used in the segment-level feature extraction approach. These facts show that emotion recognition by modeling segment-level features is an important research issue. In addition, the speech database with emotions from real-world experiences is not available and current speech databases are subjectively evaluated by other persons. The method of constructing an emotional speech database elicited by real experience and a more reliable evaluation method are also important research issues.

2. Research purpose and targets

The research purpose is to propose a new emotion recognition method using speech signals. The research targets are to propose a new classification approach of emotion recognition using speech, construct emotion database with natural emotional speech elicited by real experiences and propose a new emotional speech data evaluation method using EEG signals.

3. Thesis structure and contents

In this thesis, a new emotion recognition method using speech signals is proposed. The performance has been evaluated by using an originally designed Japanese emotional speech database elicited by real experiences, which contains speech data selected by self-assessment and EEG-assessment. This thesis composes of six chapters. Chapter 1 introduces the research background, the importance of automatic emotion recognition using speech, and clarifies research purpose and targets. Chapter 5 concludes this research and Chapter 6 demonstrates the future perspectives. The contents of Chapter 2, 3, and 4 are summarized as follows.

In Chapter 2, emotional valence and arousal recognition methods using EEG are proposed. As EEG signals are known as one of the most reliable signals for emotion recognition, they are used as references in this study for selecting high quality emotional speech data. For achieving high accuracy for emotion recognition using EEG signals, a novel feature extraction strategy in the time-frequency domain using discrete wavelet transform (DWT) is proposed. For taking into consideration useful information obtained from different frequency bands of brain activity along the scalp, a new set of features named the cross-level wavelet feature group (CLWF) is introduced. Raw EEG signals are decomposed to seven levels using DWT according to EEG bands from delta to gamma bands. Different statistical parameters are clustered and visualized using principal component analysis (PCA) for studying the effectiveness for demonstrating affective states. The standard deviation shows more capability for capturing emotional information among the statistical parameters of mean, standard deviation, skewness, and kurtosis. For searching an optimal combination of wavelet coefficients from different EEG electrodes, GA is designed to select one from eight decomposed wavelet coefficients. This procedure can be considered as feature selection with prior knowledge guidance. In addition, a reduced EEG set using less number of electrodes are studied and discussed, the conclusions from analytical analysis are supported by previous research indicating the frontal area brain activities can better demonstrate emotions. In summary, high accuracy (more than 90%) EEG-based valence and arousal recognition method is proposed; cross-level wavelet features extracted from EEG signals are robust for valence and arousal recognition.

In Chapter 3, an emotional speech database is constructed. In order to evaluate the robustness of proposed speech emotion recognition method with emotional speech data elicited by real experiences, an experimental procedure for collecting speech signals under different emotional states

is designed. The experiments consisted of two parts, which were an online survey and onsite experiments. The Internet survey was designed to collect materials representing participants' real emotional experiences. After collecting the materials for emotion elicitation, the onsite experiments were arranged to collect the speech and physiological signals. The participants recalled their emotional experiences and described them during the experiment, and the coordinator asked questions and made small talk about the same emotions with prior knowledge from the survey that they had previously collected. Finally, signals from more than 50 people were included in the database. Self-assessments are proceeded after emotional experience recalling of each emotion and EEG-assessments are conducted after the onsite experiments. Four emotions are defined on the arousal-valence space so that emotions during speaking can be identified using EEG signals by arousal and valence recognition method developed in Chapter 2. In summary, a high quality emotional speech database is constructed; experimental procedure for collecting emotional speech data elicited by real experience is designed and EEG assessment is proposed to select high quality data.

In Chapter 4, a purely segment-level speech emotion recognition method is proposed. Previous speech emotion recognition schemes are reviewed, and the reasons of adopting segment-level approach are presented. Previous existing segmentation approaches are reviewed and discussed. After addressing the quantitative analysis of various analytical schemes related to segment-level speech emotion recognition, an automatic approach for selecting a number of the most representative samples was proposed in order to improve the classifier generalization ability. Inspired from the ATIR segmentation method which has an advantages of getting a smaller and fixed number of segments from an utterance, three segmentation strategies of entropy-based ATIR (eATIR), mutual information-based ATIR (miATIR), and correlation coefficients based ATIR (crATIR) were proposed,. A more efficient way is adopted for finding more informative segments by minimizing the amount of dependency between feature vectors. Fixed-length segments are constructed in this study at selected positions based on the designed indexes. Therefore, a part of the segments is used in the analysis. This strategy overcomes the randomness of deleting speech data using other segmentation approaches such as getting segments with absolute time interval at relative times according to the length of utterances. A emotion recognition model is established using these segment-level speech frames at selected positions. The decision for determining the emotion of an utterance is based on the prediction of its segments by applying the majority vote. Two and four labels emotion recognition has been carried out both on a 50-person emotional speech database. In summary, purely segment level speech emotion recognition method is proposed; segment selection based on correlation coefficients is essential for feature extraction and majority voting is proposed for deciding the emotion of an utterance. The performance of purely segment features is improved by more than 20% compared to the approach adopting global features of utterances.

4. Conclusions

A new emotion recognition method using speech signals is proposed based on segment-level approach, which totally abandons utterance-level features. This approach can largely increase the accuracy compared to current utterance-level emotion recognition by more than 20% evaluated using 50 persons' emotional speech data. The robustness of applying this approach on natural emotional speech signals is further confirmed by an emotion speech database elicited by real experiences proposed in this study. Natural speech signals are collected under basic human emotions and the database is evaluated by self-assessment and using EEG signals as references. The proposed cross-level wavelet features largely increase emotional valence and arousal recognition accuracy using EEG signals to more than 90%, where previous research accuracy is around 40%-70%.

5. Future perspectives

A very interesting potential application area is emotion strength analysis by using segment-level speech emotion recognition. Preliminary testing has been carried out using the International Affective Picture System (IAPS) for evoking emotions with different strengths including weak and strong pleasure/displeasure to get data for preliminary testing. By the statistically analysis of segment-level labels of all the samples, the proposed method can indeed reflect the strengths of emotions in utterance clusters for a number of spoken phrases or short sentences over a short period of time. Another potential application can be short time emotion variance monitoring based on segment-level emotion recognition. High accurate EEG based emotion recognition method can be added on current brain-computer interfaces (BCI) including user emotion monitoring during driving, gaming, controlling robotic arms, etc.