報告番号	※甲　　第　　号

主 論 文 の 要 旨

論文題目　Speech Emotion Recognition in Real Environments using Characteristics of Emotional Expression and Perception
（感情の表出・知覚特性を利用した実環境における音声感情認識）

氏　　名　安藤　厚志

論 文 内 容 の 要 旨

Speech is one of the most basic and important forms of human communication. It consists of three components: linguistic, non-linguistic (those cannot be controlled consciously such as gender, age, and emotions), and para-linguistic information (those can be controlled consciously like intentions and attitudes). In order to realize speech communication between humans and machines, a great deal of research has been conducted on linguistic information recognition, i.e. automatic speech recognition. More recently, research on the recognition of non-/para-linguistic information has attracted much attention to achieve more natural communication that understands speech as well as humans do. This thesis focuses on the recognition of speaker's emotion, one of the important factors of non-linguistic information.

Emotion plays an important role in speech communication. All the speaking behaviors such as linguistic contents and attitudes are influenced by emotions. Therefore, speech emotion recognition is essential for understanding speech communications. There are a lot of practical applications such as supporting agents or "voice of the customer" analysis, human-like spoken dialogue systems that empathize with the speaker's emotions, and human psychological state detection like driver's irritability.

Although many emotion recognition studies have been conducted, there are two difficulties in real environments. First, the expression of emotion is extremely complex and diverse. Speaker's emotion is expressed by any or a combination of prosodic, linguistic, and dialogic features. For example, negative feelings can appear in a low tone of voice, negative words, long pauses, and little backchannels. It is very difficult to capture all of these characteristics to recognize emotions. Second, emotions are subjective information that is strongly influenced by the perceiver (listener). The criterion of emotion perception may differ from listener to listener. For example, some listeners perceive the speaker to be happy about an utterance, while others perceive the speaker to be in a normal state. However, the conventional emotion recognition studies ignore this listener dependency and just estimate the majority-voted emotion of multiple listeners, which results in mismatching between outputs of automatic emotion recognition system and user's feelings in real applications.

To achieve highly accurate emotion recognition in real environments, this thesis performs two-step researches. The first step is the detection of particular emotions in a real but limited sound environment. The constraints of the target emotions and environments mitigate the diversity of emotional cues, the first problem, which brings the recognition to a practical level of accuracy. Furthermore, these constraints decrease the differences in the emotion perceptions between listeners. The second step is the recognition of the wider range of emotions in a diverse real-world environment. This step aims to solve the second

problem since the differences in perceived emotions among listeners will be larger.

The task of the first step is customer satisfaction estimation in contact center calls. It can be applied to automatic agent evaluations. Two levels of customer satisfactions, turn-level and call-level, are estimated in this task. The main problem is that it is difficult to capture complex emotion expressions that appear in prosodic, linguistic, and dialogic cues. To solve this problem, a novel customer satisfaction estimation framework named a hierarchical multi-task learning model is proposed. The key idea of the proposed method is to leverage two characteristics of a customer's emotional expression. First, the satisfaction degrees of customers depend on the context. For example, dissatisfied emotional states tends to continue several turns while satisfied are not. Second, call-level and turn-level satisfaction results are closely related to each other. Calls in which the call-level satisfaction is satisfied tend to show satisfied turns in the middle and end of the call. The proposed model learns these characteristics of emotion expressions to employ Recurrent Neural Networks (RNNs) and multi-task learning of two-level estimation tasks. Experimental results on two datasets, simulated and real calls, show that the proposed method significantly improves the estimation accuracy of both call-/turn-level customer satisfaction estimations compared to the conventional method.

The second step tackles four-class basic emotion classification in natural speech. One of the applications is emotion-aware dialogue control in spoken dialog systems. The problem is though emotion perception varies with the listener in natural speech, most of the conventional methods ignore this individuality and just model the majority decision of multiple listeners. This thesis presents a new emotion recognition framework that models the emotion perception of individual listeners. The proposed method named as a listener-dependent model can estimate not only the perceived emotion of each listener but also the majority decision. It is inspired by the domain adaptation in deep learning, which has achieved great success in speech processing. Emotion classification experiments on two datasets demonstrate that the proposed method significantly improves the accuracy of listener-dependent emotion recognition.

These two studies demonstrate that there are certain trends in the expression and perception of emotional information, and that emotion recognition performance in real environments can be improved to utilize these trends. This thesis contributes to the advancement of the use of emotion recognition in real environments and the realization of natural communication between humans and machines.