氏　　　　　　　名	REDA EL-SAID MOHAMED EL-SAYED EL-BAROUGY
学　位　の　種　類	博士(情報科学)
学　位　記　番　号	博情第 284 号
学 位 授 与 年 月 日	平成 25 年 9 月 24 日
論　　文　　題　　目	A Study on Constructing an Automatic Speech Emotion Recognition System based on a Three-Layer Model for Human Perception Perception(ヒトの知覚を模擬する三層構造モデルを用いた感情音声認識システムの構築に関する研究)
論 文 審 査 委 員	主査　赤木　正人　　北陸先端科学技術大学院大学　教授
	党　建武　　同　教授
	鵜木　祐史　　同　准教授
	田中　宏和　　同　准教授
	北村　達也　　甲南大学　教授

論文の内容の要旨

The voice is an extraordinary human instrument. Every time we speak, our voice reveals our gender, age, culture background, level of education, native birth, emotional state, and our relationship with the person spoken to. All these clues are contained in even small speech segment, and other people can read our voices with remarkable accuracy. When we speak, we “encode” important information about ourselves; when we listen to others, we can “decode'” important information about them. One of the goals of human-computer interaction (HCI) is the improvement of the user experience, trying to make this interaction closer to human-human communication. Inclusion of speech emotion recognition was one of the key points to include “perception” to multi-media devices. This improved their user interfaces. However, the analysis of emotional states by the study of the implicit channel of communication (i.e. the recognition of not only what is said but also how it is said) may improve HCI making these applications more usable and friendly.

We can communicate using speech from which various information can be perceived. Emotion is an especial element that does not depend on the content of the utterance and is useful in communications that reflects the speaker's intention. Most previous techniques for automatic speech emotion recognition focus only on the classification of emotional states as discrete categories such as happy, sad, angry, fearful, surprised, and disgusted. However, emotions are usually gradually change from weak to high degree. Therefore, an automatic speech emotion recognition system should be able to detect the degree or the level of the emotional state form the voice. Hence, in this study we adopt the dimensional descriptions of human emotion, where emotional states are

estimated as a point in a three-dimensional space. These dimensions are suitable for representing the gradient nature of emotional state.

This research is concerned with the automatic speech emotion recognition system based on the dimensional model. In this model, human emotional state is represented as a point in a space consists of three dimensions: valence, activation, and dominance. Valence is used to describe emotion in terms of positive and negative assessments (e.g. happy and encouraging have positive-valence whereas angry and sad have negative-valence). Activation is used to define emotion in terms of arousal or excitation (e.g. happy and angry have positive-activation while sad and bored have negative-activation). The dominance dimension indicating the degree of weakness or strength of an expression, this dimension used to distinguish between the close neighborhood of anger and fear in the valence-activation space. The input for the automatic system are acoustic features extracted from speech signal and the output are the estimated values of valence, activation, and dominance. These estimated values for the three dimensions not only identify the emotional state but also the degree of the emotional state such as “low happy”, “happy”, “very happy”.

Conventional speech emotion recognition methods using the dimensional approach are mainly focused on investigating the relationship between acoustic features and emotion dimensions as a two-layer model, i.e. acoustic feature layer and emotion dimension layer. However, using this model has the following problems: (i) we do not know what acoustic features are related to each emotion dimension (ii) the acoustic features that correlate to the valence dimension are less numerous, less strong, and more inconsistent, and (iii) the values of emotion dimensions are difficult to estimate precisely only on the basis of acoustic information. Due to these limitations, values of the valence dimension have been particularly difficult to predict by using the acoustic features directly.

The ultimate goal of our work is to improve the conventional dimensional method in order to precisely predict values of the valence dimension as well as improve prediction of those of the activation and dominance. To achieve this goal, we construct an automatic speech emotion recognition system by adopting a three-layer model for human perception described by Scherer (Scherer, 1978) and developed by Huang and Akagi (Huang and Akagi, 2008). It was assumed that, a listener perceives the acoustic features and internally represented them as a smaller perception e.g. adjectives describing emotional voice such as Bright, Dark, Fast, and Slow. These smaller percepts or adjectives are finally used to judge the emotional state of the speaker.

In this thesis, the proposed idea to improve automatic speech emotion recognition system can be done by imitating the process of human perception for emotional state from the speech signal. The conventional two-layer model has limited ability to find the most relevant acoustic features for each emotion dimension, especially valence, or to improve the prediction of emotion dimensions from acoustic features. To overcome these limitations, this study proposes a three-layer model to improve the estimating values of emotion dimensions from acoustic features. Our proposed model consists of three layers: emotion dimensions (valence, activation, and dominance) constitute the top layer, semantic primitives the middle layer, and acoustic features the bottom layer. A semantic primitive layer is added between the two conventional layers acoustic features and emotion dimensions.

We first, assume that the acoustic features that are highly correlated with semantic primitives will have a large impact for predicting values of emotion dimensions, especially for valence. This assumption can guide the selection of new acoustic features with better discrimination in the most difficult dimension. The second assumption is that human can judge the expressive content of a voice even without the understanding of one language, such as emotional state of the speaker from different language. Using the second assumption, we investigate the universality of the proposed speech emotion recognition system to detect the emotional state cross-lingually.

To sum up, the aims of this work is to investigate the following assumptions: (1) Selecting acoustic features based on the proposed three-layer model of human perception will help us to find the most related acoustic features for each emotion dimensions. (2) Using these selected acoustic features, as inputs to an automatic emotion recognition system will improve the accuracy of all emotion dimensions especially valence. (3) In addition, we investigate whether there are acoustic features that allow us to estimate the emotional state from the voice of a person no matter what language he/she speaks. We are interesting to build a global automatic emotion recognition system, which have the ability to detect the emotional state regardless of language.

Therefore, the method we adopt to construct our speech emotion recognition system includes the following steps: first, we proposed a new acoustic feature selection algorithm to select the most relevant acoustic features for each emotion dimension by using a top-down method. Then, we build a perceptual three-layer model for each emotion dimension using a top-down method, one emotion dimension in the top layer, the highly correlated semantic primitive to this dimension in the middle layer, in the bottom layer the highly correlated acoustic feature to the highly correlated semantic primitives in the middle layer. Finally, a button-up method was used to estimate values of emotion dimensions from acoustic features by firstly, using fuzzy inference system (FIS) to estimate the degree of each semantic primitive from acoustic features, and then using another FIS to estimate values of emotion dimension from the estimated degrees of semantic primitives.

The proposed emotion recognition system was validated using two different languages (Japanese and German) in two different cases (speaker-dependent and multi-speaker). Firstly, the system was implemented for each language individually to investigate whether the system can be applied for any language. Secondly, the common acoustic features between the two languages are used to validate the second assumption.

The experimental results reveal that by using the proposed features selection algorithm for the two databases, we found many related acoustic features for each emotion dimension. The estimation accuracy for emotion dimensions is improved using the selected features comparing with all features. Moreover, the three-layer model can be applied for the two-different language databases with similar performance. The most important result is that the proposed three-layered model outperforms the conventional two-layered model. The speaker-dependent vs. multi-speaker emotion estimation was tested; it was found that the performance of speaker-dependent is better that multi-speaker. Finally, the estimated values of emotion dimensions are mapped into the given emotion categories using a Gaussian Mixture Model classifier for the Japanese and German databases. For the Japanese database, an overall recognition rate was up to 94% using emotion dimensions. For the German database, the recognition rate was up to 95.5% for speaker-dependent tasks.

In order to investigate whether the automatic system can detect the emotion dimensions for one language by training the system using different language. The proposed speech emotion recognition system was trained using Japanese language and tested using German language and vice versa. It was found that the cross-language emotion recognition system could estimate emotion dimensions with small error comparing the estimation results from a system trained using the native language.

The results indicated that the three-layer system shows an internal structure of human perception clearly and has the recognition accuracy better than that of the two-layer system. In a sense of imitating the perception mechanism of humans, the constructed system provides a more effective emotion recognition system compared with the conventional methods.

論文審査の結果の要旨

音声には大きく分けて言語情報（何を話しているか）と非言語情報（感情、個人性等）が含まれる。音声コミュニケーションではこれら両方が送受されている。このため、人－人の対話解析に基づいて人－機械のインターフェースを構築しようとする場合、言語情報（音声認識）だけではなく、話し手の感情がどのように変化しているかという情報（感情認識）は重要な要素となる。現在、感情認識の研究では、感情をカテゴリととらえ、従来型のパターン認識技術、すなわち音声認識・文字認識等で使用されてきた「入力を各感情カテゴリに振り分ける技術」（カテゴリ判別器）が用いられている。しかし、この方法が感情認識本来の目的を達成しているかどうか甚だ疑問である。なぜならば、人は、同じ感情（たとえば怒り）でも「少し怒っている」あるいは「かなり怒っている」というように感情の程度まで知覚している。また、一つの発話文から「怒っているけど悲しそうだ」などの複数の感情を知覚する。このため、機械による感情認識においても、複数の感情を同時にその程度までを含めて認識するシステムを構築する必要がある。

本論文では、従来の感情認識システムが感情をカテゴリとして捉えていたのとは異なり、「感情空間は多数の感情基本因子ベクトルによって張られる連続した多次元空間である」として捉える、すなわち、感情空間を基本因子ベクトル Arousal － Valence － Dominance の合成ベクトルとして表現するという新しい発想のもと、音声中の感情知覚モデルを感情音声認識に適用し、感情が複数含まれる音声からそれぞれの感情の程度までを推定する手法を提案している。具体的には、

(1) 感情空間の再定義：表現豊かな音声の知覚特性を扱う目的で、三階層構造感情知覚モデル

（音響特徴量、印象表現語群、感情基本因子）を提案した。感情基本因子としては、“怒り”、“恐れ”、“喜び”などのラベルではなく、感情の印象を表現できる Activation － Evaluation － Dominanceの3次元を採用した。この結果、感情を複数の基本因子ベクトルの合成ベクトルとしてより簡単に表現できるようになり、認識システムの構築が容易となった。
(2) 音響特徴の抽出および知覚モデルの改良：多数の音響特徴から感情基本因子 Arousal － Valence － Dominance の程度の推定を行うために、感情にかかわる適切な音響特徴を選択する手法について検討した。これらの検討により、特に従来難しいとされていた Valence について精度の良い推定が行えるようになり、Arousal － Valence － Dominance の 3 つの基本因子ベクトルの合成ベクトルとして感情の推定が行える土台ができた。
(3) 感情空間へのマッピングモデルの評価：提案している三層構造感情知覚モデルを用いて、推定された感情基本因子ベクトル Arousal － Valence － Dominance の組み合わせにより感情空間へのマッピングを行う手法について検討を行った。感情空間へのマッピングについて、聴取実験から得られたヒトの応答特性と比較した結果、従来手法よりもヒトの応答特性の模擬性能は高くなっており、三層構造感情知覚モデルと FIS を組み合わせた場合に、最も性能が高いことが分かった。
(4) 感情認識実験：感情音声認識パイロットシステムの構築を行い、感情認識実験の精度を議論した。日本語およびドイツ語の感情音声に対して、本手法と従来手法であるGMMを用いた手法を適用した場合の認識精度を比較した結果、本手法が認識率で大きく優れていることが確認できた。

以上のように、本研究は新しい概念のもとで音声中の感情認識システムを構築し、高い認識精度を実現したものであり、学術的に貢献するところが大きい。よって博士（情報科学）の学位論文として十分価値あるものと認めた。