氏　　名	XUE, Yawen			
学位の種類	博士(情報科学)			
学位記番号	博情第 396 号			
学位授与年月日	平成 30 年 9 月 21 日			
論文題目	A Study on rule-based emotional voice conversion with degree continuously controllable in dimensional space using the inverse three-layered model			
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	教授
		吉高　淳夫	同	准教授
		能勢　隆	東北大学大学院工学研究科	准教授

論文の内容の要旨

In terms of human-computer interaction (HCI), synthesized speech has burgeoned at a rapid rate in recent years to fulfill demands for daily speech communication. Natural sounding synthetic speech with only linguistic information is currently used in modern applications such as text to speech systems (TTS), navigation systems, robotic assistants, story teller systems and speech to speech translation systems (S2ST). Information conveyed by speech should be summarized through linguistic information, paralinguistic information as well as nonlinguistic information. Synthesized speech with only linguistic information cannot encompass all of these factors.

Therefore, emotional synthesized speech that allows communication of nonlinguistic information is increasingly required. Emotions, ranging from an underlying emotional state to full-blown emotions, contribute substantially to the acoustic manifestation of the spoken language. In order to incorporate emotions to neutral TTS synthesized speech, an emotional voice conversion system which can convert the neutral speech to emotional one is necessary to cope with.

Previous prevalent methods concerned with emotional voice conversion systems mainly used statistical approach such as Gaussian mixture models, deep neural network, or neural network method. They mainly utilized some simple discrete labels such as joy, sad, anger to represent emotion. However, humans have ability to express mild emotion such as a little bit joy or very joy which is a continuum of non-extreme states. This means statistical approaches need large parallel linguistic database of affective voices with a continuum of non-extreme affective states which is a costly and impossible problem. Little attention is paid on controlling the emotional degree in a continuous scale in previous studies.

A rule-based voice conversion technique which can obtain variation tendencies of acoustic features with a

limited database is utilized in this research.

When modeling continuous controllable degrees of emotional synthetic speech, two primary problems are firstly needed to be considered. The first one is how to describe emotions and another problem is how to model emotion perception process of human beings.

In the literature, there are many descriptive systems for emotions. The most straightforward description is the utilization of emotion-denoting words or category labels called emotion category. Emotions in daily speech communication are highly diverse. Many human-machine dialogues need machines to express mild and non-extreme emotional states. Therefore, an emotion dimensional approach which satisfies the requirement to express a range from low-intensity to high intensity states is appropriate for representing a continuum of non-extreme emotional states for controlling the degree of emotion. In this research, two dimensions arousal (synonymous to activation and activity) and valence (synonymous to evaluation) are used for emotional speech conversion based on the database we have.

Another problem related to emotion conversion is modeling the process of expression and perception of emotion by human beings. Many researchers based their theory and research on a modified version of the Brunswik's functional lens model of perception. Huang and Akagi proposed a three-layered model for expressive speech perception with emotion (listener attributions) at the top layer, semantic primtives (proximal percepts) at the middle layer, and acoustic feature (distal indicators) at the bottom layer. They assumed that humans perceive emotion not directly from acoustic features, but semantic primitives, such as fast, bright, and so on also play important roles. The three-layered model had already been applied by many researchers in the emotion recognition area. In this research, we assume that the human production of emotion follows the opposite direction of human perception. This means the encoding process of the speaker is the inverse process of the decoding of the listener. In that case, an inverse three-layered model is employed as the structure between emotion and acoustic feature.

The related acoustic features to each dimension are investigated as applied to emotional speech synthesis. Subjects are asked to evaluate the synthesized speech, the specific acoustic features of which, such as F0, have been replaced by the F0 from the emotional speech but leaving the other acoustic features of the neutral speech. We concluded that both the F0 trajectory and spectral sequence are important to emotion conversion. The power envelope and duration show little influence on the valence axes. In this research, we focused more on the prosody-related features such as duration, F0 and power envelope.

In order to control the emotion degree in dimensional space using the inverse three-layered model, an emotion conversion system was proposed with two inputs (positions in dimensional space and neutral speech) and two

steps (rule extraction and rule application). In the first step, the rules between acoustic feature variations of neutral and emotional ones can be extracted using a fuzzy inference system. The inverse three-layered model is set as the structure between emotion dimension and acoustic features with emotion dimension as the bottom layer, semantic primitive layer in the middle and acoustic features layer at the top. The second step is to apply the rule-based voice conversion method to modify the acoustic features of neutral speech to emotional ones following the rules extracted from the first step. It is widely understood that emotion is conveyed by means of a number of prosodic parameters such as voice quality and speech rate as well as fundamental frequency. In this step, some essential prosody features such as duration, F0 contour and power envelope are parameterized by an interpolation method, Fujisaki model and target prediction model. Then the modified acoustic features are synthesized using STRAIGHT. Perceptual evaluation results in V-A space show that the synthesized speech of joyful, sad and cold angry emotion can be perceived well, including the category and the degree, although the perceived degree is decreased compared to the desired values. For hot anger emotion, since the spectral modification was not conducted, the synthesized speech of hot anger is perceived as a joyful emotion.

Commonalities and differences of human perception for perceiving emotions in speech among different languages in dimensional space have been investigated in Han et al., 2016. Results show that human perception for different languages is identical in dimensional space. According to this result, we assume that, given the same direction in dimensional space, we can convert the neutral voices in multiple languages to emotional ones with the same impression of emotion. It means that the emotion conversion system could work for other languages even if it is trained with a databases in one language. We try to convert neutral speech in two different languages, English and Chinese using an emotion conversion system trained with Japanese database. We find that all converted voices can convey the same impression as Japanese voices. On the case, we can make a conclusion that given the same direction in dimensional space, the synthesized speech among multiple language can convey the same impression of emotion. In a word, the Japanese emotion conversion system is compatible to other languages.

In conclusion, this research proposed a method for emotional voice conversion with degree continuously controllable using dimensional representation following human emotion production mechanism. Perception results show that the synthesized stimuli can be perceived with the same tendency as intended in dimension space except hot anger. Neutral voices in other languages were directly inputted into the system without training, and perception results show that the conversion system built in one language is capable for other languages without training. The emotional navigation systems, robotic assistants and S2ST system will bring an intelligent HCI and enormous promotion in human life quality. As this system enables to convert the input neutral speech from any target speaker in any language without training, it can reduce an amount of cost and make the emotional TTS applicable. And this will give a big progress in the field of emotional voice conversion. The emotional navigation systems, robotic assistants and S2ST system will bring an intelligent

human computer interface HCI and enormous promotion in human life quality.

Keywords: Emotional voice conversion, rule-based speech synthesis, emotion dimension, three-layered model, Fujisaki F0 model, target prediction model.

論文審査の結果の要旨

本論文は，感情の強さを制御しながら平静音声を様々な感情音声に変換できるシステムの構築に関する研究報告である。

音声には豊かな情報が含まれている。人 - 人の音声コミュニケーションでは，言語情報だけではなく，音声に含まれるその他の情報（パラ言語／非言語情報）も送受され，コミュニケーションを円滑にしている。しかし，現有の音声翻訳システム（speech to speech translation system）であるとか，ロボットとの音声コミュニケーションでは，未だに言語情報の送受のみが行われており，話者の状況すべてを送受できているとは言い難い。特に現有の感情音声合成システムでは, HMM などの統計的手法あるいは編集合成手法を用いて，喜び・怒り・悲しみなどの特定の感情カテゴリ中の音声を合成している。しかしながら，ヒトは日々の音声コミュニケーションの中で感情の状態とか強さを自在に変化させている。このため，特定の感情カテゴリ内の音声を合成するだけでは，自然な音声コミュニケーションを実現する上で十分ではない。

本論文では，感情の状態とか強さを任意に表現可能な合成システムを構築するために，次の手法を提案した。(1) 感情の状態とか強さを自在に連続的に記述できる感情空間表現法を採用し合成音で表現したい感情を感情空間中の座標として記述した。(2) 感情知覚の三層構造モデルを用いて，感情空間中の座標から平静音声と所望の感情音声の音響特徴の差分を推定した。(3) この差分から得られる変形規則に基づいて，平静音声を所望の感情音声に変換する手法を実現した。そして，(4) 元音声を自由にかつ容易に変換するために，本論文では韻律の動的特徴をパラメータ化する二つの手法（F0 包絡制御のための藤崎モデルおよびパワー包絡制御のためのターゲット予測モデル）を採用した。

被験者を用いた評価実験の結果，意図した感情は十分な精度での感情強度・自然性を伴って知覚されることが示された。本システムによりカテゴリ内の感情音声を合成できるだけでなく，同一のカテゴリ内においても感情空間での感情強度を精度よく制御できることが分かった。本論文で提案したシステムは，パラ言語／非言語情報を音声に付加することができるシステムであり，あらゆる場面で豊かな情報を送受できる可能性がある。

以上のように，本研究は新しい概念のもとで，平静音声を様々な感情音声に変換するための手法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。

以上、本論文は、ホストゲスト化学を応用した革新的な医薬応用展開に資するシミュレー

ション科学基礎分野での重要な課題の 1 つを洗い出し、大規模シミュレーションを駆使した系統的な研究調査により新たな知見を提供した業績として、学術的に貢献するところを認め、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判断した。