氏　　名	ZHU Zhi
学位の種類	博士(情報科学)
学位記番号	博情第 381 号
学位授与年月日	平成 30 年 3 月 23 日
論文題目	Contribution of temporal modulation cues on the perception of speaker individuality and vocal emotion
論文審査委員	主査　鵜木祐史　北陸先端科学技術大学院大学　教授 赤木正人　同　教授 党建武　同　教授 荒木友希子　金沢大学　准教授 古川茂人　日本電信電話株式会社コミュニケーション科学基礎研究所　部長

論文の内容の要旨

The ability of speech communication should be the biggest difference between human and any other animal. Human speech carries not only the language message (linguistic information) but also nonlinguistic information such as speaker individuality and vocal emotion. The speaker individuality information can be defined as the information that is used by human to distinguish such a specific speaker from any other person. The vocal emotion information can be defined as the information that is used by human to identify the emotion state of speaker from speech. Both speaker individuality and vocal emotion play an important role in the speech communication of our daily life. Understanding the mechanism of how human can perceive nonlinguistic information from speech should be very important for the clarification of the mechanism of speech perception. However, the perceptual process of speaker individuality and vocal emotion is still not fully clarified at present.

Previous studies about the perception of nonlinguistic information were always based on the source- filter theory from the viewpoint of speech production. The basic reason is that nonlinguistic information can be thought to be derived from human vocal organs. The contributions of typical acoustic features conveyed in speech, such as F0, spectral envelope, intensity, and speech rates, were investigated. However, it was found that such typical acoustic features have difficulty to account for the human response from cochlear-implant (CI) listeners. A probable reason is that, for CI listeners, the temporal modulation cues provided by the temporal envelope are used as primary cues, however, the typical acoustic features can not represent the features of the temporal envelope well. The temporal modulation cues provided by the temporal envelope are also considered to be important for perceiving nonlinguistic information.

Why the temporal modulation cues provided by temporal envelope of speech should be important and needed to be clarified. At first, from the viewpoint of auditory, temporal envelope plays an important role in human

auditory system. The signal processing in peripheral auditory system can be roughly modeled as band-pass filtering (auditory filterbank) and envelope extracting (inner-hair cell model). The sound signal is first divided into several narrow band signals by auditory filterbank. Then the temporal envelope of each band is extracted as the mechanism of inner hair cells. Furthermore, it is suggested that human auditory system carries out a kind of modulation frequency analysis on the temporal envelope that can be modeled as a modulation filterbank. The auditory system should analysis the modulation frequency components at the early stage close to the periphery. Therefore, the temporal modulation cues provided by temporal envelope may contribute the perception of nonlinguistic information.

For speech perception, the temporal envelope has also been proved to be an important cue in the perception of linguistic information. Studies using noise-vocoded speech (NVS) demonstrated that human can perceive linguistic information with using the temporal envelope as a primary cue. NVS can be generated by dividing speech signal into several narrow bands and replacing the carriers in each narrow band with band-limited noise. The spectral cues provided will be poorer and poorer with less number of channels. It is shown that NVS with only four bands is sufficient to achieve good vowel, consonant, and sentence recognition. Furthermore, previous studies also showed that the low modulation frequency components of temporal envelope should contribute to the perception of linguistic information. If the temporal modulation cues are so that important to speech perception, they should also contribute to the perception of nonlinguistic information.

The clarification of the contribution of temporal modulation cues to the perception of speaker individuality and vocal emotion should be important to the development of CI devices. CI devices provide the temporal envelope information as a primary cue; however, the temporal fine structure information is not effectively encoded. As the poor spectral cue, CI listeners have problem with identifying the speaker or the emotion from only speech. It is necessary to clarify the contribution of temporal modulation cues on the perception of nonlinguistic information to optimize the CI device and improve the performance of speaker and vocal-emotion recognition of CI listeners and also for the clarification of the perceptual process of nonlinguistic information. Furthermore, the clarification of the contribution of temporal modulation cues to the perception of speaker individuality and vocal emotion will also deepen our understanding of the temporal modulation information and speech perception. To clarify the contribution of temporal modulation cues, an analysis method purely based on the modulation frequency analysis mechanism of the auditory system is necessary.

The ultimate research goal of the present study is to clarify the contribution of temporal modulation cues to the perception of speaker individuality and vocal emotion. To reach that goal, at first, the role of temporal envelope and modulation frequency information in speaker and vocal emotion recognition was investigated to confirm whether temporal modulation cues actually contribute to the perception of speaker individuality and vocal emotion. Speaker and vocal-emotion recognition experiments using NVS were carried out to investigate the effects of different temporal and spectral resolutions of NVS on the perception of speaker individuality and vocal-emotion. The spectral and temporal modulation cues will be reduced when the spectral and temporal resolution decrease.

For spectral cue, the speaker distinction performance was not sensitive to the spectral resolution, at least in the limited set of stimuli in the present study. For vocal-emotion recognition, the spectral resolution was important for the recognition of only neutral, joy, and cold anger NVS, but not sadness or hot anger NVS. For temporal modulation cues, the results showed that the recognition rates were significantly decreased with lower upper limit of modulation frequency for both speaker and vocal emotion. In the other word, it was more difficult to recognize the speaker or vocal emotion from NVS if the temporal modulation cues provided were reduced. Therefore, it was confirmed that the temporal modulation cues contribute to the perception of speaker individuality and vocal emotion. Compared to the perception of linguistic information, the temporal modulation cues provided by higher modulation frequency bands are suggested to be important for the perception of speaker individuality and vocal emotion.

At the next step, the relationship between the modulation spectral features and the perceptual data obtained from speaker and vocal-emotion recognition experiments was analyzed to clarify the exactly contribution of temporal modulation cues on the perception of speaker individuality and vocal-emotion. The modulation spectral features were extracted from the modulation spectrogram of speech data. The modulation spectrogram was calculated by the process of auditory filterbank, temporal envelope extraction and modulation filterbank. The correlation between the discriminability index of modulation spectral features and the perceptual data was calculated to demonstrate the relationship between modulation spectral features and the perception of speaker individuality and vocal-emotion.

For speaker individuality, there were positive correlations between the modulation spectral features and the perceptual data of speaker distinction experiment. For vocal emotion, similar results were also obtained, however, the correlations were roughly higher than that of speaker distinction experiments. The results showed that the modulation spectral features were useful to account for the perceptual data of speaker and vocal-emotion recognition experiments using NVS. Therefore, modulation spectral features were suggested to be important cues contribute to the perception of speaker individuality and vocal emotion.

Finally, applications of the temporal modulation information in simulating CI listeners' response and vocal-emotion conversion of NVS were discussed. At first, the feasibility of using NVS to simulate CI listeners' response in vocal emotion recognition was investigated by carried out vocal-emotion recognition experiments using both NVS and original emotional speech with NH and CI listeners. The results showed that the vocal-emotion recognition paradigm using NVS can be used to investigate vocal emotion recognition by CI listeners. Furthermore, it was also suggested that the modulation spectral features can also be used to account the performance of CI listeners in the vocal-emotion recognition.

Effect of the modification of modulation spectrogram on the vocal-emotion recognition was then investigated. A method based on a linear prediction (LP) scheme was proposed to modify the modulation spectrogram and its features of neutral speech to match that of emotional speech. The temporal envelopes were modulation-filtered by using IIR filters to modify the modulation spectrum from neutral to emotional speech. The IIR filters were derived from the relation of modulation characteristics of neutral and vocal emotions on a

LP scheme. On the acoustic frequency domain, the average amplitude of the temporal envelope was corrected using the ratio of the average amplitude between neutral and emotional speech. Finally, a vocal-emotion recognition experiment using NVS generated by the converted temporal envelope was carried out. The results showed that the modulation spectrogram of neutral speech can be successfully converted to that of emotional speech by the proposed method. The results of the evaluation experiment confirmed the feasibility of vocal emotion conversion on the modulation spectrogram for NVS.

In conclusion, the fact that the temporal modulation cues contribute to the perception of speaker individuality and vocal emotion was confirmed by the speaker and vocal-emotion recognition experiments using NVS. Furthermore, the investigation of modulation spectral features demonstrated that there were high correlations between modulation spectral features and the perceptual data obtained from speaker and vocal-emotion recognition experiments. Therefore, the modulation spectral features could be important cues contribute to the speaker and vocal-emotion recognition with NVS. These results further proved that the temporal modulation cues play an important role in the perception speaker individuality and vocal-emotion.

Keywords: Speech perception, speaker individuality, vocal emotion, temporal cue, noise-vocoded speech, modulation spectral feature

論文審査の結果の要旨

音声には言語情報の他に非言語情報が含まれている。特に、非言語情報の一つである個人性や感情は音声コミュニケーションで重要な役割を果たしている。個人性や感情に関係する音声の音響特徴は、音声生成からのアプローチにより，徐々に明らかにされつつあるが，それらの音響特徴と個人性・感情認識といった聴知覚との間の関係性については未だ明らかになっていない。これらの関係を明らかにできれば，ヒトの音声知覚のメカニズムを解明できるだけでなく、機械による話者認識や感情認識にも応用できるはずである。音声の言語知覚、特に音声の明瞭性・了解性に関しては、音声信号の振幅包絡線情報に含まれる振幅変調成分に重要な手がかりがあることがわかっている。特に、変調周波数 4 Hz 付近の変調成分が重要であることがわかっている。もし振幅包絡線の変調成分に非言語情報も含まれているのであれば、音声の言語・非言語知覚と聴知覚との関係性について、振幅変調を切り口として一貫した検討を行うことが可能である。

本研究では、音声の個人性及び感情情報の知覚に寄与する特徴が、音声の振幅包絡線（変調成分）に含まれるのかどうか、さらに、含まれるとすればそれらはどのような変調成分の特徴であるかを検討した。まず、ヒトの聴覚末梢系での信号処理を模擬するために、聴覚フィルタバンク及び振幅包絡線検出処理に相当する変調フィルタバンクを構築し、雑音駆動型音声合成を利用した振幅包絡線情報ベースの知覚実験処理系を整備した。次に、こ

の処理体系に基づき、音声の振幅包絡線情報の変化に関して個人性ならびに感情知覚への影響を調査した。その結果、振幅包絡線情報が非言語知覚に影響を与えること、特に変調周波数 4〜16 Hz の変調成分が重要であることを明らかにした。最後に、これらの変調成分に対する高次統計量の分析から、音声の個人性や感情の知覚に重要な変調スペクトルの特徴を明らかにした。これらの成果は、人工内耳装用者による非言語情報の知覚実験の結果を説明できるだけでなく、雑音駆動型音声合成における非言語情報の変換・強調処理にも活用することができることから、人工内耳用音声信号処理の機能向上に大きく貢献できる。

以上、本論文は、雑音駆動型音声合成を利用した知覚実験から音声の振幅包絡線情報に非言語情報が含まれることと、その知覚に重要な特徴を明らかにしたことから、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。