氏名	郭　太陽		
学位の種類	博士（情報科学）		
学位記番号	博情第 564 号		
学位授与年月日	令和 7 年 9 月 24 日		
論文題目	Contributions of temporal amplitude envelope and its instantaneous modulation components on vocal-emotion perception		
論文審査委員	鵜木　祐史	北陸先端科学技術大学院大学	教授
	岡田　将吾	同	教授
	長谷川　忍	同	教授
	吉高　淳夫	同	准教授
	古川　茂人	静岡社会健康医学大学院大学	教授
	坂本　修一	東北大学	教授

論文の内容の要旨

In everyday communication, speech plays a vital role, particularly when other modalities such as text, facial expressions, or body language are unavailable. Accurate understanding of the information conveyed by speech is essential for effective communication. According to Fujisaki, speech conveys three types of information: linguistic, non-linguistic, and paralinguistic. Among these, non-linguistic information, such as vocal emotion, often reflects the true intent of the speaker and emotional state. Therefore, correctly interpreting emotional information from speech signals is crucial in daily life.

Speech signals can be considered as a combination of temporal fine structure (TFS) and temporal amplitude envelope (TAE). TFS is primarily responsible for pitch perception and sound localization. The TAE contributes to the identification of phonemes and the recognition of consonants, vowels, and words. Numerous studies using noise-vocoded speech (NVS), synthesized by using only the TAE and white noise, have demonstrated that for normal-hearing (NH) listeners, the TAE plays a more significant role than the TFS in processing linguistic information. Especially, the TAE shows greater robustness to hearing loss and aging compared to the TFS. In addition, TAE reflects the temporal structure and modulations of speech, influencing the perception of intensity and duration of segments and pauses. Therefore, TAE contributes significantly to emotion perception.

In addition, because cochlear implants (CI) rely on the TAE as an important cue for speech perception, NVS has been widely used to simulate auditory perception of CI users. However, CI users often face challenges in vocal-emotion perception compared to NH listeners. NH listeners also face some challenges, such as the decline in emotion recognition accuracy in noisy environments. To address these issues, it is essential to identify the important components of the TAE and explore how controlling them can improve the accuracy of emotion recognition.

Previous studies have already suggested the important role of TAE in vocal-emotion perception by conducting vocal-emotion perception experiments. On the other hand, psychoacoustic

and physiological evidence in previous studies over the past decades has supported the existence of the selectivity of the modulation frequency in the auditory system, highlighting the importance of modulation processing of TAE in the auditory system. Building on this foundation, previous research has also shown that modulation spectral features of the TAE and the modulation frequency components of the TAE contribute to emotion perception.

This study focuses on temporal modulation frequency features related to stress, accent, and pauses as important modulation components of the TAE in emotion perception. These features are represented by the instantaneous modulation components (IMCs) of the TAE. Specifically, this study focuses on the IMCs of the TAE, specifically, the instantaneous modulation frequency (IMF) and instantaneous modulation amplitude (IMA). The main objective of this study is to clarify the contribution of IMCs and to investigate whether vocal-emotion perception can be controlled by manipulating IMCs. To do so, this study conducted three main experiments on vocal-emotion perception using NVS.

Experiment I aimed to clarify the important modulation frequency components of the TAE by applying a modulation filterbank to band-pass filter modulation frequency components. The results of Experiment I suggested that modulation frequency components in the 0–16 Hz band are important.

Experiment II aimed to investigate whether the dynamic components of modulation frequencies, IMCs, play an important role in vocal-emotion perception. To do so, we first verified whether the contribution of modulation frequency components alone can explain the results of vocal-emotion perception. We applied time-reversal processing to the TAE, which preserves the amplitude spectrum of the TAE modulation components in the 0–16 Hz while reversing the temporal order of TAE temporal modulation features. The results showed that the decrease in emotion recognition rates suggested that, in addition to modulation frequency components, temporal modulation features of TAE play an important role in vocal-emotion perception. We then introduced the concept of IMCs to represent temporal modulation features of TAE and used temporal stretching and/or compression processing on IMCs to simulate the time-reversal processing. The results showed that temporal processing on IMCs could simulate time-reversal processing and explain the results of vocal-emotion perception. It is suggested that IMCs play an important role in vocal-emotion perception.

Experiment III aimed to investigate whether the contribution of IMCs not only decreases but also enhances or restores emotional cues. To this end, we conducted vocal-emotion perception experiments using temporal processing that applied the inverse of the stretching and compression operations used in Experiment II. The results showed that, for certain emotions, temporal processing of IMCs enhanced vocal-emotion perception compared to the original NVS, and it also restored emotional cues that were degraded in time-reversed NVS. In other words, vocal-emotion perception can be controlled by manipulating IMCs to either enhance or reduce recognition rates. This highlights the role of IMCs in vocal-emotion perception and suggests their potential for

addressing challenges in emotional speech processing in future applications.

In addition, the three experiments described above were conducted with NH listeners using NVS. Since NVS simulates the auditory experience of CI users, it is essential to investigate whether the results of this study are generalizable to actual CI users in real-world settings. Therefore, we extended the experiments to include CI users to explore the applicability of our results in practical CI settings. We conducted additional experiments in the appendix using the same experimental settings to assess vocal-emotion perception in CI users. The results indicated that CI users exhibited recognition patterns similar to those of NH listeners. These findings suggest that the contribution of important modulation frequency components and IMCs of the TAE may also apply to vocal-emotion perception in CI users.

In conclusion, this study clarified that the IMCs of TAE in the 0–16 Hz band of the TAE contribute to vocal-emotion perception. Moreover, it showed that temporal processing of IMCs can lead to both deterioration and enhancement of emotion recognition accuracy. These results provide a more comprehensive understanding of the role of TAE modulation components in vocal-emotion perception and suggest the potential of temporal processing of IMCs to improve vocal-emotion perception in both NH listeners and CI users in future applications.

Key words: Vocal emotion perception, Temporal amplitude envelope, Instantaneous modulation components, Noise-vocoded speech, Modulation filterbank

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．音声の言語・非言語情報に関係する音響特徴は，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．一方で，聴知覚メカニズムの検討から，音声の振幅包絡線情報が言語・非言語情報に知覚に重要であることが明らかにされている．特に，言語情報（明瞭性）に関しては振幅包絡線情報に含まれる 4 Hz 付近の振幅変調成分が，非言語情報（感情認識）に関しては 4～16 Hz の振幅変調成分が重要であることが明らかにされている．これらの成果は，人工内耳装用者による言語・非言語情報の知覚に重要な知見をもたらしており，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ものと期待されている．

本研究では，感情知覚における振幅包絡線情報の重要な変調成分として，ストレス，アクセント，休止に関連する時間的変調周波数特性に焦点を当て，その認識メカニズムの解明に取り組んだ．また，人工内耳シミュレータの一つである雑音駆動音声合成を利用して，振幅包絡線情報の瞬間変調周波数成分が感情知覚にどのような関連性があるか，3 つの主要な実験を実施して検討した．まず，1 つ目の実験では，瞬時変調周波数を取り扱うための変調フィルタバンクの利用が聴取実験の結果に影響を与えないことを確認した．また，変調フィルタバンクを利用して感情知覚に重要な変調周波数帯域が 0～16 Hz であることを明らかにした．2 つ目の実験では，時間包絡線情報を時間反転したものとそうでないものの間の感情知覚結果から，変調周波数帯域が同じであっても時間変化が異なることで感情知覚に違いが生じることを明らかにした．これは瞬時変調周波数の重要性を示している．3 つ目の実験では，瞬時変調周波数成分の時間変化を振幅包絡線情報の時間伸長圧縮処理により，2

番目の実験の違いを説明できること，さらには感情知覚のエンリッチメントを実現できることを明らかにした．これらの成果は，今後の人工内耳用音声信号処理で重要な知見となる．

以上，本論文は，感情知覚に重要な特徴が音声の振幅包絡線情報の瞬時変調周波数成分にあることを明らかにした．これは新規性と独創性を持ち，その成果が国際的に定評のある学術誌に掲載されるなど学術的水準も高い．本技術は，感情知覚の包括的な理解と聴知覚メカニズムの解明の一助になり，応用範囲が広く，学術的に貢献するところも大きい．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．