| | |
|---|---|
| 氏　　名 | 李　　凱 |
| 学位の種類 | 博士（情報科学） |
| 学位記番号 | 博情第 525 号 |
| 学位授与年月日 | 令和 6 年 3 月 22 日 |
| 論文題目 | Advanced Feature Representation of Audio Signal Based on Spectral Temporal Modulation Analysis and Its Applications |
| 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 |
| | SAKTI Sakriani Watiasri　同　准教授 |
| | 吉高　淳夫　同　准教授 |
| | 岡田　将吾　同　准教授 |
| | 赤木　正人　同　名誉教授 |
| | LU Xugang　情報通信研究機構　主任研究員 |

**論文の内容の要旨**

Audio, including speech, music, machine sounds, etc., surrounds us every day. In recent years, a lot of advanced pattern recognition technologies have been proposed by using different kinds of audio and labels, greatly facilitating our human lives. For example, human and machine-synthesized speech is collected and used to perform automatic speaker verification (ASV) and fake audio detection (FAD) systems, which can be applied in authentication and access control, justice, healthcare, speech security, etc. Furthermore, the audios of different kinds of machines from the factory are collected and utilized to construct machine anomalous sound detection (ASD) systems, which can provide continuous monitoring of machine status and optimize production efficiency. For these techniques, extracting acoustic features that can capture distinguishing information serves as the foundation for achieving accurate performance.

Many acoustic features have been proposed and applied in different audio- related tasks, such as the linear prediction coefficient (LPC), the Mel-frequency cepstral coefficient (MFCC), and the constant Q cepstral coefficient (CQCC). These features are designed for general information extraction and can be used for many different tasks. For example, the MFCC and LPC are widely used as acoustic features for both automatic speech recognition (ASR) and ASV. However, there are different kinds of information, such as linguistics, individuality, and emotion, encoded in the audio signal, and different tasks need to have task- specific information (TSI) to separate out different patterns. Traditional acoustic features, such as MFCC and LPC, have problems with weak task-specific discrimination, resulting in extracted features containing a lot of redundant information or important information being filtered or smoothed out.

Spectral temporal modulation (STM) analysis in the auditory system deals with both spectral and temporal modulation of audio to perceive auditory attributes related to audio production, such as the timbre and prosody. This property enables the human ear to easily perceive and discriminate a wide range of TSI in various acoustic scenarios. Inspired by this, this study aims to investigate advanced feature representations based on STM analysis to extract more TSI for audio detection and verification tasks.

To achieve this objective, both frequency and time domain analyses are conducted to explore the importance of spectral and temporal attributes in the representation of TSI. Then, STM representations derived based on the

frequency and time analysis results are proposed for extracting more TSI. The frequency and time domain analyses can help verify the effectiveness of spectral/temporal modulation and direct better designing for the auditory models. The effectiveness of proposed feature representations is verified in ASV, FAD, and machine ASD, which cover speech, machine-synthesized speech, and non-speech. The complexity of the human auditory mechanism makes it challenging to fully understand the audio signal processing process and determine which auditory model best mimics this process.

In frequency domain analysis, this research first investigates the importance of different frequency regions for ASV, FAD, and machine ASD tasks. Specifically, we proposed two data-driven-based methods, including the F-ratio and a frequency-wise attention structure combined with a ResNet, to quantify the nonlinear combined effect of frequency components. We then designed a non-uniform subband processing strategy based on the quantification results for task-specific feature extraction. The quantification results from these three applications consistently indicated that TSI distributed in the frequency domain non-uniformly, which is quite different from traditional auditory scales, such as the equivalent rectangular bandwidth. The designed non-uniform filterbanks can capture more TSI and further improve the performance of these three applications.

In time domain analysis, this study investigates the timbre and prosody information differences in the voice represented using the jitter and shimmer features. In accordance with the statistical analysis results, the most promising prosody features were selected and incorporated with a machine-learning-based FAD system. In addition, two additional F0 estimation methods, namely YIN and IRAPT, were utilized in place of the IRAPT algorithm when extracting features. Different weights were tested to determine the optimal combination between the Mel- spectrogram and shimmer features. The experimental results indicate that both the static and dynamic shimmer features of voice can provide complementary knowledge to the traditional spectrum-based systems in the FAD task.

The STM representations simulate the complex auditory perception process in the human auditory system, capturing both spectral and temporal modulations in speech signals. The effectiveness of STM has been evaluated in speech intelligibility prediction and speech emotion recognition. Usually, the Gammatone filterbank is utilized in the extraction processes of STM representations. According to the results of frequency domain analysis, the non-uniform filterbanks are more effective in extracting TSI. Therefore, this study simulated the human auditory mechanism using the STM derived from well-designed non-uniform filterbanks and evaluated in SV, FAD, and machine ASD tasks. The experimental results indicate that the STM representations can achieve competitive performance in the FAD and machine ASD tasks, which covering synthesized speech and non-speech signals. However, in the ASV task, the current results are inconsistent with our expectations. One possible reason may be the mismatch between the proposed representations and the i-vector approach used for ASV, which are designed for short-term feature extraction. More experiments regarding to the deep-learning architectures will be conducted in the future.

**Keywords**: Spectral temporal modulation, Human auditory perception, Fake audio detection, Automatic speaker verification, Machine anomalous sounds detection.

**論文審査の結果の要旨**

近年，サイバーフィジカル空間における音声コンテンツの利用は，スマートフォンの普及やAIスピーカ等の登場とともに急激な伸びを示している．このような急激な需要拡大に対して，音声情報を安心・安全に利用するための技術革新や法整備は相当な遅れをとっており，音声コンテンツの不正利用やなりすまし，音声改ざんといった問題も招いている．そのため，サイバーフィジカル空間において，ディジタル表現された音声メディア情報を安心・安全に利用するために，音声コンテンツの真正性と話者情報の真偽性を検証する基盤技術を確立する必要がある．特に，これらの問題に対して，話者認証技術や音声なりすまし検出，ディープフェイク音声検出といった基盤技術が提案され，深層学習を利用して，その高度化が試みられている．

本研究では，上述した基盤技術を確立する上で重要となる音響特徴表現を検討し，様々な音声アプリケーションを実行する上で重要となる前処理（特徴抽出）について議論している．従来，音響特徴量としてMFCCやメル対数スペクトルなどが利用されてきたが，これらは話者識別やなりすまし検出，フェイク音声検出，異常音検出といったタスクに対し，適切であるかどうか議論されずに利用されている．本研究では，各タスクに対し，タスク固有の情報を取り扱う必要があるという立場を取りつつ，それを包括的に表現できるものとしてスペクトル変調・時間変調情報（STM）に着目している．この特徴には，韻律など音声生成に関連する情報だけでなく，音色といった聴知覚に関わる聴覚属性の情報も含まれる．特に，周波数次元の特徴としてF比から算出されるNonuniform Filterbank出力が，時間次元の特徴として振幅変調成分の他にJitter & Simmerが利用され，これらを統合した時間・周波数次元での変調情報表現を特徴づけるSTMを使用することを提案した。これらの特徴を３つのタスク（話者識別，フェイク音声検出、異常音検出）に適用し、それぞれのタスクで従来よりも有益な結果を導き，タスク固有の情報を包括的に表すSTMの重要性を示唆した。

聴覚研究の分野では，STMは聴覚末梢・中枢から上位に向けられる聴覚的な特徴表現として考えられている．本研究の結果は，STMを利用した音声の特徴表現が音声処理の様々な応用課題に有益であり，新規性と有効性が認められる．また，その特徴抽出法が聴知覚メカニズムにも強く関係することから，音全般の知覚ならびに応用課題に大きく寄与するものと考えられる．

以上，本研究は，聴知覚メカニズムに着目した音声特徴表現の重要性・有責性を示し，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．