Abstract

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 timbral 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 F_0 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.