論文の内容の要旨

Characterization and classification of speech and non-speech acoustic events
(音声・非音声特徴づけと自動分類に関する研究)

氏名 エスピ　マルケス　ミケル

（ 本文）Automatic detection of acoustic events provides information that describes human and social activities. Besides the fact that speech is usually the most informative of the signals occurring in acoustic scene, it is not the only one. Human actions, passive or active, produce sounds, and this can provide information that would be impossible to obtain otherwise (e.g. a sneeze could mean that people is feeling cold although the thermometer does not show a low temperature). Therefore, event-triggered systems are the most direct beneficiary, providing: implicit assistance to the users inside the room, context-aware and content-aware information requiring a minimum of human attention or interruptions, support for high-level analysis of the underlying acoustic scene, etc. On the other hand, the recent fast growth of available audio or audiovisual content strongly demands tools for analyzing, indexing, searching and retrieving specific contents within those databases. Moreover, allowing systems to be aware of the current acoustic scene can lead to the improvement of automatic speech recognition technologies, resulting into better informed system with further precise confidence measures for error estimation.

As the title states, the objective of this thesis is not only to classify acoustic events but also to provide an adequate characterization by means of features and models that fit better the targeted acoustic events. This contrasts with most recent developments in acoustic event detection, which have focused on robust classifiers and feature selection, yet using features that fit better for speech by definition as they were designed for speech related tasks. The thesis is driven by three main ideas. First that speech is a property-rich acoustic event, and that each of those properties allows to differentiate speech from a certain group of sounds (e.g. loudness is robust in a silent room, while pitch would allow to differentiate speech from wind), noting that none of those properties are invariantly robust.Second, in the context of non-speech acoustic event an approach similar to speech is not possible since non-speech acoustic events are not so property-rich and are very heterogeneous, so we look at human perception to provide a model that integrates acoustic properties in a knowledge-based manner. And third, with the idea of defining a language model for non-speech acoustic events, we introduced the \emph{situation} model, a framework based on topic models for acoustic events.

In voice activity detection, the observation of speech spectrum leads to the fact that speech has a specific spectral fluctuation pattern both in time and frequency. Short and long term dynamics of spectral features, although providing robustness against environmental noise, are still significantly affected by the presence of non-stationary noise, especially periodical noises. Speech signals have a specific spectral fluctuation behavior regarded as intermediate between harmonic and percussive sounds, fluctuating slow along time, and fast along frequency. We have separated speech components based on their specific spectral fluctuation behavior, and a set of features have been extracted for voice activity detection. A specific two-stage harmonic-percussive sound separation procedure has been used for speech separation and the resulting multi-feature VAD has been compared with conventional features in voice activity detection, reducing frame-wise detection error by up to 78% depending on the signal-to-noise ratio conditions and noise type.

Detection of non-speech acoustic event supporting heterogeneous sets of events face the problem of having to characterize them when they have different acoustic properties (more transient, more stationary, both, etc.). Moreover, managing large feature vectors with features characterizing different properties of the signal is always difficult.
In order to cope with we observe how humans focus on different properties depending on which kind of acoustic event are being compared (e.g. is not the same differentiating a drum from a door knock, than differentiating between speech and wind). We automated this with a modified tandem connectionist model in which we first enhance the feature information separately by properties, then we combine these resulting features, and then a decision is made upon these resulting features. This is solved by changing the traditional early integration scheme into a late integration scheme.

Finally, we introduce a novel acoustic event detection framework aiming to provide a pseudo-language model for acoustic events. Such model is not ad-hoc but flows from the assumption that recordings can not only be segmented in acoustic events (such as speech, steps, etc.), but they can also be segmented in ``situations'' on a higher abstraction level (such as people gathering, coffee breaks, presentation time}, etc. in a conference for instance). The proposed model also exploits the fact that non-speech acoustic events have very fixed spectro-temporal structures with very little variation, claiming that: first, spectro-temporal context can be characterized directly at the feature level; and second, a probabilistic model based on topic models can incorporate a pseudo-language model on top providing better classification.