Thesis Title: A Study on Audio Features for Acoustic Scene Classification

ABSTRACT

Growing demands from context-aware devices, surveillance, and archiving applications have fuelled research towards efficient extraction of useful information from environmental sounds. Assigning textual labels to audio segments based on the general characteristics of recording locations is called acoustic scene classification (ASC). This dissertation focuses on studying well-known audio feature-extraction methods and investigating their relevance for recognizing environments. The proposed framework is based on the experiments conducted on two datasets from detection and classification of acoustic scenes and events challenges' (DCASE, 2016, and 2017) ASC task. The systems are benchmarked with the baseline systems of the two challenges.

Unlike speech, environmental audio has overlapping frequency content while spanning a larger audible frequency range. Also, the signals are less structured than speech/music signals. Keeping this in mind, we investigate the spectral features all-pole group delay function (APGDF), constant-Q cepstral coefficients (CQCC), mel-frequency discrete wavelet coefficients (MFDWC), non-overlap block transform coefficients (NOBTC), and spectral centroid frequency coefficients (SCFC) from speech recognition and speaker verification research. All these features are inspired by the human audio perception in different ways. While NOBTC, MFDWC and CQCC consider the magnitude of the signal's' time-frequency representation, APGDF and SCFC extract scene discriminating information from the phase and frequency counterparts respectively. We present a detailed analysis on the suitability of the frame-level statistics of these features for environmental audio processing and optimize their parameters accordingly. The proposed framework employs support vector machine as the classifier.

Because of the diverse nature of audio scenes, a single feature-classifier pair may not efficiently differentiate environments. A collective decision from all participating systems of the DCASE challenges was found to surpass the accuracy obtained by each system. We evaluate data, decision, and sensor fusion strategies for combining the information extracted from the features mentioned above and a set of short-term time and frequency features for an improved ASC performance.

For many ASC applications, a general estimate of the surroundings (e.g., indoor or outdoor) might be enough, while some others might require to work only in locations belonging to a particular type of environment. In this thesis, we propose a two-level hierarchical framework for ASC. At the first level, texture features extracted from time-frequency representation of the signals are used to generate the coarse labels. The system then employs a fusion of the features as mentioned earlier for second-level classification to produce finer labels.

Keywords: Constant-Q transform, Environmental audio, Hierarchical classification, Information fusion, Local binary pattern, Mel-scaled cepstral features, Support vector machine, Texture features, Wavelet transform.