Automated Classification of Music Genre, Sound Objects, and Speech by Machine Learning

Leigh M. Smith,[*1] Stephen T. Pope[#2], Jay Leboeuf,[*3] Steve Tjoa[*4]

[*]*iZotope Inc., USA*

[#]*HeavenEverywhere.com, USA*

[1]lsmith@izotope.com,[2]stephen@heaveneverywhere.com, [3]jleboeuf@izotope.com, [4]stjoa@izotope.com

ABSTRACT

Background

Listeners demonstrate remarkable skill in identifying the source or principle character of individual sounds with very little (e.g short duration) stimuli. This demands efficient coding of the spectro-temporal behaviour of sound for classification, identification and interpretation. Recent developments in machine learning provide a means to evaluate the representative power of features derived from the sound signal and hence suggest those coding methods that human listeners may use.

Aims

A software system, MediaMined (MediaMined 2012), is described for the efficient analysis and classification of auditory signals. This system has been applied to the tasks of musical instrument identification, classifying musical genre, distinguishing between music and speech, and detection of the gender of human speakers. For each of these tasks, the same algorithm is applied, consisting of low-level signal analysis, statistical processing and perceptual modeling for feature extraction, and then supervised learning of sound classes. Given a ground truth dataset of audio examples, textual descriptive classification labels are then produced. Such labels are suitable for use in automating content interpretation (auditioning) and content retrieval, mixing and signal processing.

Main Contribution

A multidimensional feature vector is calculated from statistical and perceptual processing of low level signal analysis in the spectral and temporal domains. Machine learning techniques such as support vector machines are applied to produce classification labels given a selected taxonomy. The system is evaluated on large annotated ground truth datasets (n > 30000) and demonstrates success rates (F-measures) greater than 70% correct retrieval, depending on the task. Issues arising from labeling and balancing training sets will be discussed.

Implications

The performance of classification of audio using machine learning methods demonstrates the relative contribution of bottom-up signal derived features and data oriented classification processes to human cognition. Such demonstrations then sharpen the question as to the contribution of top-down, expectation based processes in human auditory cognition.

Keywords

Sound object classification, signal processing, machine learning, music genre

REFERENCES

MediaMined (2012). Retrieved June 15th, 2012 from http://www.mediamined.com