

Audio Signal Classification: An Overview

David Gerhard
School of Computing Science
Simon Fraser University
Burnaby, BC
V5A 1S6
email: dbg@cs.sfu.ca

Abstract

Audio signal classification consists of extracting physical and perceptual features from a sound, and of using these features to identify into which of a set of classes the sound is most likely to fit. The feature extraction and classification algorithms used can be quite diverse depending on the classification domain of the application. This paper presents an overview of the current state of the audio signal classification research literature.

1 Introduction

Audio signal classification (ASC) is a field of research that has historically been explored in a few very concentrated areas, with less work done on the general problem. The individual pieces have traditionally been speech recognition and related problems, music transcription [Moo77] [Pis86], and recently speech/music discrimination [Sau96] [SS97]. Other problems in the field have been researched as well, but the main direction has been toward speech and music applications.

The major exception that has recently appeared is the multimedia database [FG94] [WBKW96]. Searching algorithms that find a piece of multimedia based on the sound must be robust and must consider much more than simply music or speech, and for that reason an algorithm that sorts sounds into appropriate categories must be able to deal with a much larger range of sounds that a speech recognizer would. This is not to say that the problem is any harder—a speech recognizer only expects speech, but must distinguish between all possible phonemes in the applicable language, as well as recognizing slurs between phonemes, word boundaries, and sometimes must even glean meaning from prosodic features such as pitch, emphasis and timing.

1.1 Classification Problems and Applications

Speech recognition is perhaps the classic ASC problem [RJ93]. Incoming sound must be classified by the phoneme being voiced at any time, and then the phonemes must be combined to form the most likely words, phrases and sentences that make sense. Speech recognition is only recently beginning to be a useful tool instead of a frustrating exercise, but it is still a long way from the dream of speaker-independent continuous recognition of non-grammatical speech in a noisy environment. There is much ASC work to be done.

Music recognition and transcription is another computing application of ASC. Here, The sound to be classified is music, the possible classes are musical notes, and the output is usually a musical score. Systems have been developed to solve this problem with one or two musical lines [Moo77], but when the sound to be transcribed contains a full orchestra, the problem is considerably harder. ASC would speed up the creation and use of melody databases which could be accessed by direct human input, in the form of humming, whistling or singing at the computer.

A more general ASC system could differentiate between speech and music, and could be used to assign a sound to a particular transcription system. If a sound were classified as speech, the speech recognizer would be employed, and if were music, the music transcriber would be called into service. In a system such as this, the speech recognizer and music transcriber could be optimized to expect only the appropriate input, simplifying each and improving the robustness of the whole system.

More publicly consumable applications of ASC include television and radio advertisement identification, for muting or VCR pausing, and radio music style recognition that would receive a command from a user, such as “Play me some country!” and

scan the available radio channels for music fitting the query. A telephone on-hold indicator could be developed to recognize when the piped-in music has stopped and a person is on the line.

ASC could be applied to equalization, the boosting or attenuating of particular frequencies in a sound to make it more intelligible or enjoyable. Because this process has traditionally been done by hand, a finite number of equalization filter settings are often employed in an audio application. Equalization filters are currently used in hearing aids as well as in public address systems, where the user must change the setting by hand depending on the input. An automatic equalization system could detect the current environment and apply an appropriate filter setting.

Each of these possible applications represent a separate ASC problem which must be examined and researched individually. The quest for the universal classifier is probably as futile in the audio field as it is in any other classification field at this point.

2 Features

The first step in any classification problem is to identify the features that will be used to classify the data. Feature extraction is a form of data reduction, and the choice of feature set can make or break a classification system. For that reason it is sometimes left to the brute power of an algorithm to decide which features are the most important. For more on automatic selection of features, see [Sch96] [DH73].

2.1 Physical and Perceptual Features

The features typically used in ASC can be divided into physical and perceptual categories. *Physical features* are properties that correspond to physical quantities, such as fundamental frequency (F_0), energy, zero-crossing rate (ZCR) and modulation rate. *Perceptual features* are properties that correspond to the way humans perceive sound. These include pitch, loudness, timbre and rhythm. Some features are related, such as pitch with F_0 and energy with loudness, but it is important to recognize that these features are not identical. Pitch is related to the log of frequency, but pure sinusoids sound sharp, and very low frequencies sound flat, compared to a purely logarithmic relationship [CWE94]. For a feature to be truly perceptual, there must be some perceptual model in the extractor used to measure the feature.

As this paper is meant to be an overview, only two features will be presented in detail along with

the more general discussion of feature extraction and analysis.

2.2 Fundamental Frequency

The *fundamental frequency*, or F_0 , of a signal is the lowest frequency at which the signal repeats, and is only relevant for periodic or pseudo-periodic signals. It is clear that extracting the F_0 from a signal will only make sense if the signal is periodic. Because of this, F_0 detectors can be used as periodicity detectors - if the F_0 extracted makes sense for the rest of the signal, then the signal is considered to be periodic. If the F_0 appears to be randomly varying or is detected as zero, then the signal is considered to be non-periodic.

F_0 can be used as a classification feature in many ways. If the F_0 changes in discrete jumps and stays on particular values, it is a good indicator of music. If a F_0 pattern repeats regularly but within the pattern the F_0 tends to sweep across values, the sound could be a birdcall. Most of the information that is obtained from examining the F_0 of a sound comes from watching how the sound changes over time.

2.3 Zero Crossing Rate (ZCR)

As the name implies, ZCR is a measure of how often the sound signal crosses from positive to negative or vice-versa. At first investigation, it might seem that ZCR would be a good technique for finding F_0 , the thought being that a sinusoid will cross the zero line twice per cycle and hence ZCR should be exactly $2F_0$. It was soon made clear that there are problems with this measure of F_0 [Roa96] If the signal is spectrally rich, then it might cross the zero line more than twice per cycle. It also became clear that ZCR is useful for extracting features other than F_0 , such as spectral content, voicedness and noise content [Ked86]. Many recent classification systems employ various statistics taken from the ZCR measure [SS97] [Sau96]. ZCR is a popular feature because it is very easy to measure— analog ZCR meters detect the change in polarity of the voltage from a sound input line without a single cycle of DSP.

2.4 Feature Duration

It is often important to examine how the feature values change over time. The instantaneous F_0 of a signal is instructive, but how the F_0 changes is often much more useful. It can be used to investigate the constancy of the F_0 , and the duration of each sound event that has frequency, as well as how much of the total signal is taken up with periodic sound. These sub-features can be applied to many other features. For example, a common feature used in

speech classification systems is whether the sound is voiced or unvoiced. If the sound is periodic, it is most likely voiced, and if it is non-periodic with strong high-frequency components, it is likely to be unvoiced. The ratio between voiced and unvoiced segments in the sound can be used in a speech/song decision, because a piece of singing is more likely to have long chunks of voiced sound punctuated by short breaks of unvoiced sound.

Another property based on feature duration is the perceptual property of rhythm. When a piece of sound is considered rhythmic, it often means that it repeats in some way, on a time scale much longer than that required to generate a frequency. Periodicity in feature values is a good indicator of rhythmicity, and this requires examining the sound at a much longer scale than most feature extractors use. Since rhythmic information, along with other types of information, can occur at many different scales, it is often useful to examine the signal at more than one resolution at a time. Techniques that do this are called *multiresolution* techniques.

2.5 Feature Clustering

Some classifications can be determined from a single feature, but most are confirmed by examining several features at once. Algorithms that do this statistically, called clustering algorithms, make use of many pieces of data. Each piece of data, called a *case* in the clustering literature, corresponds to an observation of a sound, and the features extracted from that observation are called *parameters*. Clustering algorithms work by examining a large number of cases and finding groups of cases with similar parameters. These groups are called clusters, and are considered to belong to the same category in the classification. Once the clusters have been discovered, a representative case is chosen for each cluster, usually corresponding to the center of each cluster, and new cases are classified depending on the proximity to the representative cases. More detailed discussions of clustering techniques are presented in [Sch96] and [DH73].

3 Conclusion

Pattern Classification is a traditional and well studied problem, and ASC is a specific instance of that problem which has not been studied in as much detail. In itself, ASC is a broad research area which can be divided into many sub-problems. ASC has many potential applications and many open research problems.

References

- [CWE94] Stanley Coren, Lawrence M. Ward, and James T. Enns. *Sensation and Perception*. Harcourt Brace College Publishers, Toronto, 1994.
- [DH73] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley and Sons, Toronto, 1973.
- [FG94] Bernhard Feiten and Stefan Günzel. Automatic indexing of a sound database using self-organizing neural nets. *Computer Music Journal*, 18(3):53–65, Fall 1994.
- [Ked86] Benjamin Kedem. Spectral analysis and discrimination by zero-crossings. *Proceedings of the IEEE*, 74(11):1477–1493, November 1986.
- [Moo77] James A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, pages 32–38, November 1977.
- [Pis86] Martin Piszczalski. *A Computational Model for Music Transcription*. PhD thesis, University of Stanford, 1986.
- [RJ93] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, 1993.
- [Roa96] Curtis Roads. *The Computer Music Tutorial*. MIT Press, Cambridge, 1996.
- [Sau96] John Saunders. Real-time discrimination of broadcast speech/music. In *International Conference on Acoustics, Speech and Signal Processing*, pages 993–996. IEEE, 1996.
- [Sch96] Jürgen Schürmann. *Pattern Classification*. John Wiley and Sons, Toronto, 1996.
- [SS97] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *International Conference on Acoustics, Speech and Signal Processing*, volume II, pages 1331–1334. IEEE, 1997.
- [WBKW96] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton. Content-based classification, search and retrieval of audio. *IEEE MultiMedia*, pages 27–37, Fall 1996.