Machine Recognition of Sound Sources

Carol P. Jaeger and Charles A. Laszlo
Department of Electrical and Computer Engineering
University of British Columbia, 2356 Main Mall, Vancouver, B.C., V6T 1Z4

Introduction

In this paper an approach for machine-based classification of various types of sounds that may be present in an "everyday" acoustic environment is described. The classification process is part of a larger project that includes the identification of the presence of a sound source, classification of the source, and localization of the sound source relative to the microphone(s) with which it was measured. We have called the classification system the "Sound Class Framework" (SCF). The SCF and examples of signal processing techniques developed based on the SCF will be the focus of this paper.

Background

Understanding speech in noise is a common problem for the hard of hearing listener. A common method for improving the speech to noise ratio is to filter the signal acquired by a microphone to emphasize the bandwidth associated with speech. While this can result in an improvement in many cases, it is still a common occurrence that the spectral properties of the speech and noise sources present in a given situation overlap. This overlap limits the amount of noise reduction that can occur using spectral filtering alone. The approach we are taking in this project is to use spatial filtering to isolate a sound source of interest before transmitting it to the listener. In an earlier project [1], we found that spatial filtering was an effective way to isolate a sound source, and that with relatively simple signal processing techniques speech sources could be located and acquired in about 1.5 seconds. However, we also found that with the signal processing used, which was designed around the spectral and temporal qualities of speech only, the system was easily confused by multiple speech and noise sources. In order to achieve good spatial filtering, it is first necessary to identify the nature and position of all sound sources in a given environment relative to the recording microphone(s). It is only then that an informed decision can be made as to which sound to focus on.

The Sound Class Framework

The Sound Class Framework is a system that we have devised for the purpose of grouping sounds by their spectral and temporal characteristics. The SCF was developed to serve as an aid to the development of efficient signal processing techniques for the automatic identification and localization of sound sources. In this section we will first describe the basic structure of the SCF, and will then explain how it is used in the development of the signal processing techniques in this project.

There are four major categories in the SCF. They are: stationary-continuous (SC), stationary-intermittent (SI), nonstationary-continuous (NC), and nonstationary-intermittent (NI). Figure 1 shows the basic structure of the SCF. In the definition of these four categories, the distinction between stationary and nonstationary refers to the spectral properties of a sound and the distinction between continuous and intermittent refers to the temporal properties of a sound. Within these four categories, there is a further subdivision of sounds into more familiar groups (eg. speech, music, alarms, mechanical noise, etc.). In the SCF, speech and music would be classified as NI because the spectral distribution and signal intensity both change with time. Alarm noises such as a telephone ringer would be classified as SI because the spectral content is fixed for the duration of the ring but the ringing sound is heard for only short-lived intervals. It should be emphasized that in this project we have restricted the groups in the SCF to sounds likely found in an "everyday" acoustic environment such as the average home or office.

The goal of using the SCF as a basis for the signal processing in this project is to find just enough information about a particular sound source to be able to localize it and assign a priority to it in as short amount of time as possible. The way in which different types of sounds are grouped in the SCF allows the categorization of the sounds, such as "this is speech" or "this is an alarm", without the need to determine the exact details of the acoustical signal. We won't know what is being said, and won't be able to tell the difference between a telephone ringer and a microwave oven beeper, but it is our contention that we can successfully prioritize sound sources without knowing this level of detail. The priorities assigned to each type of sound source might vary from user to user, but a logical protocol being used as our default is as follows:

- under normal circumstances speech would be ranked as the highest priority sound source taking precedence over music and noise from machinery
- in the absence of a speech-only source the priority would shift to music or music and speech combinations (such as might come from a television)
- alarm and warning signals would take precedence over all other sources, speech or otherwise.

With the SCF in place, the next step in the project is the development of signal processing techniques that quickly and successfully identify sounds and automatically assign them to the correct group in the SCF.

Signal Processing Techniques

The signal processing for this project is divided into four parts. The first part is the identification and characterization of the sound sources present. The second part is the localization of the identified sounds relative to the recording microphone(s). The third is to create

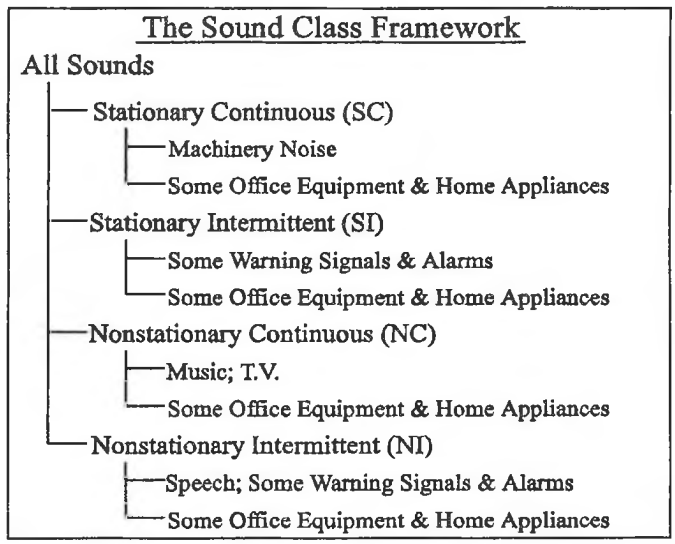

Figure 1: Structure of the Sound Class Framework. Note that some groups fit under more than one main category.

and/or update our "acoustic map" of the room and select a sound for amplification. The fourth is the control of an adaptive directional amplification system (eg. a mechanically controlled directional microphone, a phased array of microphones, or a series of fixed directional microphones arranged to accommodate all possible directions). In this paper we will describe the methodology of the first part of the signal processing, the identification and characterization of sounds, and give some examples.

We have taken an approach to signal processing that we are calling a "mixed transform technique". In essence, this means that we are using a variety of signal processing tools to achieve the desired goal of efficient localization and classification of sound sources. Sounds are identified based on unique "features" that they exhibit in the results of the algorithms applied. A library of features for different sounds has been developed and is being expanded on an ongoing basis.

The signal processing is structured in an iterative fashion, where initially fast and simple algorithms are used to get preliminary information about the acoustic environment. This is called the overview stage, and pending the results of the overview further algorithms are executed as necessary -- the detail stage. This iterative process is designed to speed up processing so that only as many algorithms as necessary are executed in order to isolate sound features.

Sound samples are recorded at a sampling rate of 16 kHz, and are divided into segments of 8192 samples long. There are three algorithms used in the overview stage: amplitude thresholding, a Discrete Wavelet Transform (DWT) on the full segment, and a Fast Fourier transform (FFT) on a 512 point subset of the segment extracted from the region of peak signal intensity. Amplitude thresholding is performed on each segment, and if no portion of the segment exceeds a predetermined threshold, it is assumed that there are no sound sources present and no further algorithms are applied to that segment. For segments that exceed the amplitude threshold, the DWT and FFT are computed.

The wavelet transform has recently become very popular in 1- and 2-dimensional signal analysis (many good texts and review articles are available on the subject, eg. [2]), and we have found a number of cases where they provide a new insight into our work.

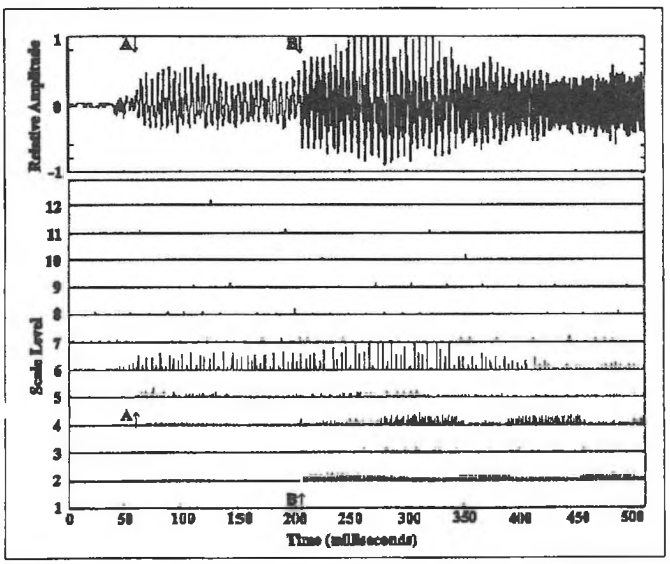

Figure 2: Time series (top) and DWT scalogram (bottom) for three overlapping sounds: the word 'test' (starting at **A**), a phone ringer (starting at **B**), and a fan (for the duration). In the scalogram the highest frequencies are represented in scale level 1.

Because the elemental building block in wavelet decomposition is localized in space and time, it seemed an obvious choice for use in sound source localization. However, we have found that the scalogram – the wavelet equivalent of a spectrogram – is also an efficient way to map a variety of different sound groups into different parts of a 2-dimensional mapping. The DWT of an 8192 point array gives 12 scale levels, each representing a different range of frequencies. The 3 or 4 levels representing the highest frequencies tend to contain information from alarm and warning signals and from music. The 5 levels representing the lowest frequencies tend to contain information from mechanical sources such as fans. The remaining levels in the mid-range of the spectrum tend to contain information from speech sources. Warning signals such as telephone ringers, alarm clocks, and other such devices have a distinctive pattern in the DWT. They generally have uniform amplitudes and repeatable patterns. Speech also has a distinctive pattern, though it is not as simple as that seen with the alarms. The low frequency noises from mechanical sources do not have a distinctive pattern that can be seen in the DWT scalogram. However, the presence of coefficients of significant magnitude in these low frequency scale levels suggests that further processing be done to identify their source. Figure 2 shows an example of both the time series and the scalogram for three overlapping sounds: a speech sample, a telephone ringer, and a fan. In figure 2 the start of the speech sample is marked with an 'A'. The distribution of the DWT coefficients in levels 5 and 6 are characteristic of speech. The start of the telephone ringer is marked with a 'B'. The alternating pattern of coefficients in levels 2 and 4 is characteristic of the two-tone electronic ringers used in many new telephones. The fan does not have a unique pattern in the scalogram, but the presence of coefficients in levels 8 through 12 suggest that a noise with a low frequency component such as a fan is present. It is easier to see the onset and duration of the different sounds in the scalogram than it is in the time series.

The FFT performed on the 512 point sample allows evaluation for a variety of features. The peaks in the FFT are isolated and from them the following features may be obtained: significant energy in the low frequency range, with few or no outstanding peaks, extending down to DC, suggests a mechanical noise source such as a fan or other equipment with either rotating parts or forced air; equally spaced peaks under 1 kHz with fundamental in range 80 to 200 Hz suggests a speech source; multiple peaks over 1kHz matching the frequencies of musical notes suggests music source. Different algorithms are executed on the results of the FFT based on the preliminary findings in the DWT.

Signal processing continues on each 8192 point segment in this fashion of fast overview analysis followed up by detailed analysis until it is appears that all sources present have been identified. In some cases the results of several 8192 point segments are concatenated together in order to identify patterns that extend beyond the 0.5 second duration of a single segment.

Summary

The method presented is an efficient and effective way of classifying sound sources. Further development is underway to automate the application of the source identification, as some user intervention is still required. Following completion of this task, algorithms for the localization of the sources and control of a directional amplification method will be implemented.

References

[1]. Harris, C.P, A Device to Localize Sound Sources, M.A.Sc. Thesis, University of British Columbia, Vancouver, BC, Canada, 1995.

[2]. Chan, Y.T., Wavelet Basics, Kluwer Academic Publishers, Norwell, Massachusetts, 1995.