RECOGNITION OF ENVIRONMENTAL SOUNDS

Richard S. Goldhor

AudioFile, Inc., 4 Militia Drive, Lexington, Massachusetts 02173 USA

ABSTRACT

This paper describes a preliminary study designed to answer the question, "How well can familiar environmental sounds be identified?" By "familiar" we mean sounds on which the recognition system has been previously trained. By "environmental sounds" we mean sounds generated by acoustic sources common in domestic, business, and out-of-doors environments. The results of this study indicate that well-isolated familiar sounds can be recognized with high accuracy by applying standard statistical classification procedures to feature vectors derived from two-dimensional cepstral coefficients.

INTRODUCTION

Although extensive efforts have been made to develop systems capable of recognizing such specialized acoustic sources as speech and submarines, much less effort has been directed towards systems capable of detecting, isolating, and identifying the panoply of sounds which fill our every-day acoustic environment. Some recent efforts of interest include [1], a signal understanding system applied to environmental sounds, and [2], in which LPC, VQ, and HMM techniques are applied to such sounds. We report here on the utility of some other signal processing and classification techniques that have proved effective in recognizing discrete speech, when those techniques are adapted to the problem of recognizing environmental sounds.

DATA COLLECTION

Recordings of environmental sounds were made under realistic background noise conditions in domestic and business environments. Background sound levels were typically 25 to 30 dB below signal levels. For sounds generated by what we judged to be "consistent" sound sources (such as a smoke alarm, but unlike a barking dog) each token (sample) was recorded with the microphone at a different distance or orientation from the sound source. Both near field and far field recordings were made, separately, of each sound. Table 1 shows the types of environmental sounds studied, and the number of near and far field tokens recorded for each source.

Sounds were recorded using a Bruel & Kjaer 2231 Sound Level Meter as a microphone. Recordings were made on a Technics SV-MD1 digital audio tape recorder.

TYPE OF SOUND	FAR FIELD	NEAR FIELD
Smoke alarm	5	5
Barking dog	10	0
Bouncing basketball	10	10
Handbell (cowbell)	5	5
Bouncing glass bottle	11	4
Smashing glass bottle	11	8
Window fan	5	5
Car engine	5	10
Doorbell (chimes)	5	5
Electric clapper bell	5	5
Hand drill	5	5
Exhaust fan	4	5
Mechanical phone bell	5	5
Electronic phone (1)	5	5
Electronic phone (2)	5	5
Screen door closing	5	5
Wooden door closing	5	5
Water running in sink	5	5
Water running in bathtub	5	5
Bouncing tennis ball	10	10
Vacuum cleaner motor	5	5
Violin (middle A)	5	5
Whistling tea kettle	5	5

TABLE 1: List of environmental sounds, and number of near and far field tokens of each sound recorded. In all, 268 tokens of 23 acoustic sources were recorded.

0-7803-0946-4/93 $3.00

The recordings were transferred from the DAT recorder to a computer as analog signals. They were redigitized at the computer at 16,000 samples/second using a 14-bit A/D converter similar to the type that would be used in a practical recognition device. Reference spectrograms were produced of each token, and a database of signal files (one token per file) were stored for further processing.

SIGNAL PROCESSING

The recorded tokens were segmented using an automatic technique based on per-token signal power histograms. Such histograms have proved an effective basis for endpointing words in speech recognition systems (c.f. [3]). Signal power was calculated for non-overlapping 16 msec blocks over the entire duration of each signal file. It was anticipated that each such histogram would be bimodel, with an upper peak corresponding to power levels within the signal itself, and a lower peak corresponding to typical background noise levels (in our case 25 to 30 dB below signal levels). From the histogram statistics, a threshold power level was calculated to separate portions of the signal stream containing the sound from portions containing only background noise.

This endpointing algorithm was applied to each of the signals in the database. Because each token had been stored in a separate file, we simply chose the segmentation limits as the first and last potential endpoint locations within each file. Given the high SNR of our recording conditions, this simple algorithm worked quite well.

We chose to use two dimensional cepstral coefficients for our feature vectors, because of their demonstrated effectiveness in speech recognition [4]. Cepstral coefficients were calculated for each token, as follows. A sequence of analysis frames were generated from each endpointed token file by windowing the samples using a 256-point Hanning window. The frame advancement rate was chosen to yield frames that overlapped by at least 25%, and so that the total number of frames between the signal endpoints was at least 64, and always an integer power of two.

Within each frame, a 256-point cepstral transform was calculated. Only the first 32 cepstral coefficients of each frame were retained. Cepstral transforms were calculated for two variants of the spectrum within each frame: a linear variant and one in which the frequency scale was warped using a mel frequency transformation.

The number of frames generated for each signal was then reduced to exactly 64. If necessary, this was done by averaging adjacent frames together (in sets of 2, 4, 8, etc). The result, in effect, was that each signal was represented by a [64 by 32] array of cepstral coefficients, with the 32 rows of the matrix representing frequency, and the 64 columns representing time. Note that by reducing each signal to exactly 64 frames we have eliminated the possibility of using token duration as information upon which to base our classification decision.

A 64-point real DFT was then calculated for each row in the matrix, and the first 32 points of this symmetrical transform retained. The resulting square matrix is a two dimensional cepstral representation of the input signal. Each row corresponds to a particular spectral frequency, and each column corresponds to a temporal frequency. The first row contains the DFT of the power envelope of the signal. The first column contains the DFT of the average signal spectrum. The first element of the first row contains the average signal power level. It is typical of two dimensional cepstral representations of acoustic signals generally, and certainly for our signals, that this corner element is the largest component, and that the size of the components in the first row and first column are larger than the size of interior matrix components.

FEATURE EXTRACTION

Small sets of coefficients were selected from the 1024 elements of each matrix to serve as feature vectors for use in classifying the sounds. Feature vectors ranged in size from two to 16 parameters (coefficients). The coefficients chosen were taken from the beginning of the first or second row, and the top of the first column, of the cepstral matrices. But in no case was the first element in the first row (the average signal power) selected as a classification feature.

CLASSIFICATION

Each feature vector defines a point in a multi-dimensional cepstral space where the sound represented by that vector is "located". The hope, of course, is that all the samples of each sound will cluster together in that space, and that clusters for different sounds will be well separated.

Clusters, or classes, were formed by grouping together the feature vectors for each type of sound. (For Experiments One through Three, near and far field samples were grouped together as the same class.) Class statistics (mean and variance vectors) were calculated and stored for each sound class.

Recognition accuracy was tested using a "jackknife" procedure. For each sound in the database, its feature vector was temporarily removed from its class, and the class statistics recalculated. *Thus no feature vector was ever tested against a class whose statistics include that vector. Each test vector is truly "unknown".* A maximum likelihood calculation was performed to determine the best classification for that test sound [5]. Recognition accuracy scores were calculated for each type of sound, and for each different feature vector definition.

EXPERIMENT ONE: ACCURACY vrs NUMBER OF FEATURES

In the first experiment, the relationship between recognition accuracy and the number of features in each feature vector was examined. The results are summarized in Figure 1, in the curve labeled "cepstrum: fully trained". When two cepstrum coefficients are used to classify sounds, recognition accuracy is greater than 60%. (In the absence of any information, less than 5% of tokens would be correctly classified.) When each sound is characterized by four features, classification accuracy is greater than 90%, and accuracy approaches 98% for 12 and 16 features.

EXPERIMENT TWO: USE OF MEL FREQUENCY SCALE

In the second experiment, mel cepstral coefficients were calculated for each windowed frame of the token waveform. All other processing was identical to the procedure described above. The results are shown in Figure 1, in the curve labeled "mel: fully trained". The overall relationship between recognition accuracy and feature count closely follows the results from Experiment One, but recognition accuracies are several percentage points lower for feature counts less than ten, and less than 0.5% lower for 16 features.

EXPERIMENT THREE: ACCURACY vrs. CLASS SIZE

In this experiment, only five tokens were used to calculate the statistics for each sound class. Recognition accuracies were tested for both cepstral and mel cepstral coefficients. The results are shown in Figure 1 in the curves labeled "5 token training". As might be expected, recognition accuracy is lower than in the first two experiments, where class sizes ranged from 8 to 19 tokens.

EXPERIMENT FOUR: EFFECT OF NEAR & FAR FIELD RECORDINGS

The analysis procedure for this experiment was identical to the previous three, except that class statistics for each sound class were calculated only from tokens recorded under near field (non-reverberant) conditions. The 12-parameter feature set from Experiments One through Three was used. The average number of training vectors per class was approximately the same in all trials (about five, as in Experiment Three, where typically two vectors were derived from near field tokens and three from far field tokens). Because we calculated the accuracy with which both near and far field tokens could be classified using these statistics, recognition accuracy figures could be derived (when the results of Experiment Three was included) for four conditions:

* near field test data following training on near field tokens only;
* far field test data following training on near field tokens only;
* near field test data following training on combined near and far field tokens;
* far field test data following training on combined near and far field tokens.

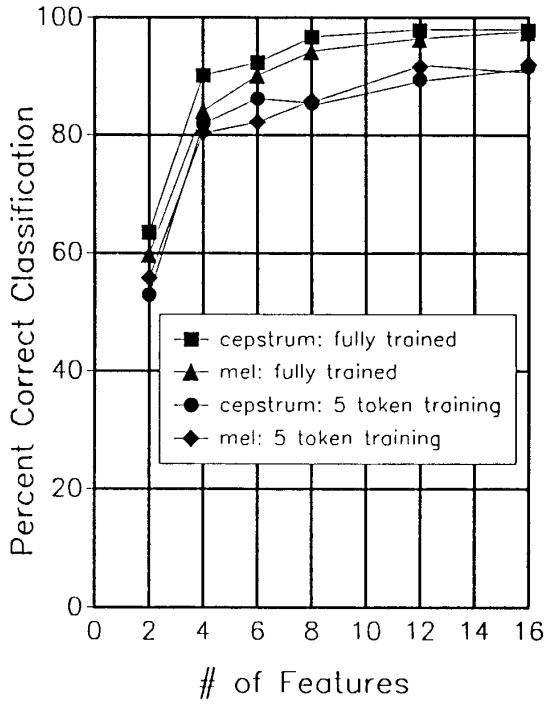

Figure 1: Percent correct classification of test tokens as a function of the number of coefficients in the feature vector representing each token.

Table 2 shows the results of this experiment. The generally low accuracy values reflect the fact that class statistics typically were calculated from only five samples. These figures suggest that reverberation does affect classification accuracy, at least for the features we have selected.

Conclusion

The simple recognition system constructed for this study effectively classifies isolated tokens from several dozen disparate sound sources. Two dimensional cepstral coefficients proved to be effective classification features. Only the cepstral coefficients representing "low frequency" spectral and temporal variations were required in order to obtain accurate classification. Mel cepstral coefficients appear to be almost as effective as linear-frequency cepstral coefficients.

The most difficult problems associated with constructing an environmental sound recognition system may prove to be dealing with reverberation, separating multiple simultaneous sound sources, and characterizing sequentially-structured sounds such as footsteps.

LITERATURE CITED

[1] E. Dorken, S.H. Nawab, E. Milios, "Knowledge-Based Signal Processing Applications", in **Symbolic and Knowledge-Based Signal Processing**, edited by A.V. Oppenheim and S.H. Nawab, Prentice Hall, 1992.

[2] J.P. Woodard, "Modeling and Classification of Natural Sounds by Product Code Hidden Markov Models", *IEEE Trans. on Signal Processing*, vol. 40, pp. 1833-1835, 1992.

[3] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, & J.G. Wilpon, "An Improved Endpoint Detector for Isolated Speech Recognition". *IEEE Trans. Acoust., Speech, Signal Processing* ASSP-29, 777-785, 1981.

[4] T. Kitamura & E. Hayahara, "Word Recognition Using a Two-Dimensional Mel-Cepstrum in Noisy Environments". paper PPP6 presented at the 2nd Joint Meeting of the ASA and ASJ, Hawaii, 1988.

[5] K. Fukunaga, **Introduction to Statistical Pattern Recognition**. New York: Academic Press, 1990.

Training Tokens	Test Tokens: Near Field	Test Tokens: Far Field
CEPSTRAL PARAMETERS		
Near Field Only	90.4	79.8
Combined Near and Far Field	82.2	96.5
MEL CEPSTRAL PARAMETERS		
Near Field Only	94.1	84.0
Combined Near and Far Field	88.3	93.6

Table 2: Percent classification accuracy resulting from training on near field tokens only, versus training on mixed near and far field tokens. Average training class size is approximately five tokens for both training conditions.