AUTOMATIC CLASSIFICATION OF ENVIRONMENTAL NOISE EVENTS BY HIDDEN MARKOV MODELS

Paul Gaunard *Corine Ginette Mubikangiey* *Christophe Couvreur* *Vincent Fontaine*

Faculté Polytechnique de Mons, 31, Boulevard Dolez, B-7000 Mons, BELGIUM
Tel: ++ 32 65 374176 - Fax: ++32 65 374129
Email: {couvreur,fontaine}@tcts.fpms.ac.be

ABSTRACT

The automatic classification of environmental noise sources from their acoustic signatures recorded at the microphone of a noise monitoring system (NMS) is an active subject of research nowadays. This paper shows how hidden Markov models (HMM's) can be used to build an environmental noise recognition system based on a time-frequency analysis of the noise signal. The performance of the proposed HMM-based approach is evaluated experimentally for the classification of five types of noise events (car, truck, moped, aircraft, train). The HMM-based approach is found to outperform previously proposed classifiers based on the average spectrum of noise event with more than 95% of correct classifications. For comparison, a classification test is performed with human listeners for the same data which shows that the best HMM-based classifier outperforms the "average" human listener who achieves only 91.8% of correct classification for the same task.

1. INTRODUCTION

The latest generation of noise monitoring systems (NMS's) is based on digital signal processing technology. They commonly implement such features as computation and storage of noise levels (L_{eq}), one-third-octave spectra, statistical indices or the detection of noise events based on thresholds. Since the computational power of signal processors keeps increasing, it is likely that NMS's will become capable of even more sophisticated treatments of the sound data they record. Consequently, research has been undertaken to develop new measurement features for inclusion in NMS's. An area of research that has started to attract much attraction recently is *automatic noise recognition* (ANR). The goal of an ANR system is the automatic —i.e., without human intervention—classification of the noise sources that are present in the acoustic environment from their recordings at the microphone of the NMS.

One particular problem in ANR is the classification of noise events such as car or truck pass-bys, aircraft fly-overs, etc. The ANR systems that have been proposed for that task rely generally on two-step process: a pre-processor converts the acoustical signal of the noise event into a set of characteristic features which are then used by a classifier to make a decision on the nature of the source of the noise event. Until now, the pre-processors that have been proposed were based on a "static" approach. That is, the noise event was reduced to a global set of characteristics which is then used to perform the classification. For instance, the average spectrum of the is a common choice. Various statistical pattern recognition techniques have been suggested for the realization of the classifier acting on that "static" representation.

In this paper, a new method for the classification of noise events based on hidden Markov models (HMM's), a technique that has been widely successful in automatic speech recognition [5, 3], is proposed. HMM-based classifiers use a "dynamic" recognition method that takes directly into account the time-frequency structure of the noise events. As will be seen, the utilization of hidden Markov models can bring significant improvement over previously proposed methodologies for the automatic recognition of noise events.

The remainder of this paper is organized as follows. In section 2, the choice of the pre-processor for an ANR system based on HMM's is discussed. Application oh HMM's to ANR is discussed in section 3. Experimental results obtained for the classification of five types of environmental noise events are presented in section 4 together with results of human listeners for the same task. Conclusions are drawn in section 5.

2. PRE-PROCESSING

For the classifier to act directly on the time-frequency structure of the signal, the pre-processor must convert the raw acoustic signal sampled at the microphone into a time-frequency representation. Such time-frequency representation can be obtained by splitting the signal into T (consecutive or possibly overlapping) short frames and compute a set of features characteristic of the spectrum for each frame. The output of the pre-processor will then be a series of spectral components $\mathbf{x} = (x_1, x_2, \ldots, x_T)$, where x_t is a set of features representative of the spectrum corresponding to the t-th frame of signal. For example, if a one-third-octave filter bank is used and short-time L_{eq}'s are computed in d frequency bands, x_t can be the d-dimensional vector formed from the d one-third-octave levels for the t-th integration interval of the L_{eq}'s. In this case, the frame length corresponds to the integration length for the L_{eq}'s.

Instead of using a filter bank, other types of spectral analysis can be used on the signal frames. In section 4, LPC (Linear Prediction Coding) cepstral analysis will be used [3].

Both the filter-bank method and LPC-cepstrum method of spectral analysis convert the original acoustic signal into a sequence of continuous-valued vectors $x_t \in \mathbb{R}^d$. This sequence of continuous-valued vectors can be converted into a sequence of discrete symbols by a technique called *vector quantization* (VQ) [4]. VQ allows the utilization of discrete HMM's.

Christophe Couvreur is a Research Assistant of the Belgian National Fund for Scientific Research (F.N.R.S.). He is also currently a Visiting Scholar with the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign.

3. APPLICATION OF HMM'S TO ANR

As the theory of HMM's is widely described in the literature, we invite the reader who is not familiar with hidden Markov modeling to refer to standard tutorials such as the ones available in [5] or [3].

In this paper, five specific types of noise event sources are considered: cars, trucks, mopeds, aircraft, and trains. Because of their transient nature, these types of noise events are well suited to be modeled by left-right HMM's. Several issues involved in the design of a HMM classifier for this environmental noise event recognition application are now discussed.

First, a spectral analysis pre-processor must be selected for the classifier and its parameters must be chosen. For the LPC-cepstral analysis pre-processor, the parameters are: the analysis frame length, the analysis frame shift, the order p of the LPC model, the number of cepstral coefficients, etc. For the filter-bank analysis pre-processor, a practical choice would be to use the one-third-octave or octave filter-banks with computation of short-time L_{eq} commonly provided by standard sound level meters.

Second, it must be decided if a vector quantization step is incorporated between the spectral analyzer pre-processor and the hidden Markov model classifier. If VQ is used, it is necessary to decide on a codebook size and a distance measure. Finally, the type of HMM's that will be used must be selected and their parameters (number of states, transition probability matrix structure, etc.) must be chosen.

Once the type of HMM and the type of pre-processor have been chosen, taking into account the external constraints, it is still necessary to find the parameter set that will yield the best performance. This can only be done with a combination of trial-and-error experiments and engineering experience, possibly guided by some physical understanding of the acoustical phenomena modeled. It is also possible to use the rules-of-thumb for the design of HMM-based classifiers which are used in speech recognition community.

4. EXPERIMENTAL RESULTS

In this section, experimental results obtained for the classification of environmental noise events with hidden Markov models are presented. Five types of noise event sources are considered: cars, trucks, mopeds, aircraft, and trains. The noise event recordings used for the training and the evaluation of the HMM's are extracted from the MADRAS database. The STRUT software provides the implementation of HMM algorithms.

4.1. The MADRAS Database

The MADRAS database of environmental noise sources has been constructed for the MADRAS project which has been partially funded by the European Community and involves several research partners in various European countries [2]. The aim of the MADRAS project (Methods for Automatic Detection and Recognition of Acoustic Sources) is to develop new noise monitoring instruments with the ability to automatically identify and quantify, in real time, the various acoustic sources which make up a given acoustic environment. The MADRAS database includes high quality recordings of various types of common environmental noise sources such as trains, cars, trucks, delivery vans, motorcycles, mopeds, aircraft, chain saws, lawnmoyers, industrial plants, etc. Several instances of each type of source are provided. The recording conditions of each noise source are documented.

For the classification experiments that will be presented here, only five types of noise recordings available in MADRAS were used: cars, trucks, mopeds, aircraft, and trains recordings.

4.2. The STRUT Software

The practical implementation of HMM algorithms is not a trivial programming task. Fortunately, software tools are available that can greatly help the realization of HMM-based classification systems. Our application of hidden Markov modeling techniques to environmental noise event recognition relies on the Speech Training and Recognition Unified Tool (STRUT) developed in the Circuit Theory and Signal Processing (TCTS-Multitel) Laboratory of Faculté Polytechnique de Mons to conduct research on speech recognition [6]. STRUT is a software toolbox that consists of many small "independent" pieces of code running on Unix (SUN, HP, Linux) and Windows workstations. Each small program implements a specific step in the speech recognition process: signal pre-processing (extraction of spectral features), vector quantization, Viterbi decoding, probability evaluation, maximum-likelihood training, classification, etc. The small programs communicate by exchanging files, through Unix pipes or through Unix sockets.

4.3. Classification Results

The HMM-based classifiers were trained on a set of noise event recordings extracted from the MADRAS database and the classification performance was evaluated on a distinct set of recordings also extracted from the MADRAS database. The partition of the events of a given type between training and test set was random. The training set contained 141 noise event recordings: 45 "car" events, 33 "truck" events, 28 "moped" events, 14 "aircraft" events and 21 "train" events. The test set contained 43 noise event recordings: 14 "car" events, 11 "truck" events, 9 "moped" events, 4 "aircraft" events and 5 "train" events. Testing the performance on only 43 samples means that the recognition rate estimates will not be very reliable, but it was not possible to use a larger testing set (or training set) because of the limited size of the MADRAS database.

In the first experiment, the pre-processor was the standard LPC-cepstral pre-processor of speech recognition. The performance of a classifier for the values of the pre-processor parameters commonly used in speech recognition applications was evaluated. The values of the parameters are:

- analysis frame length $w = 30$ ms
- analysis frame shift (overlapping factor) set to one third of w,
- order of auto-regressive analysis $p = 10$
- number of cepstral coefficients equal to 12
- VQ codebook size $L = 256$.

A three-state HMM was used ($M = 3$). The HMM also included three "silence" states at its beginning and at its end. This classifier correctly recognized 91% (39/43) of the test samples.

In the next series of experiments, the pre-processor was still the LPC-cepstral pre-processor but, this time, parameters of the pre-processor (w, p) and the number of states of the HMM's (M) were varied. Only the most significant results will be presented here. A more complete description of the results obtained can be found in [7]. Figure 1 shows the influence of the analysis frame length w on the recognition rate for single-state, three-state, and five-state

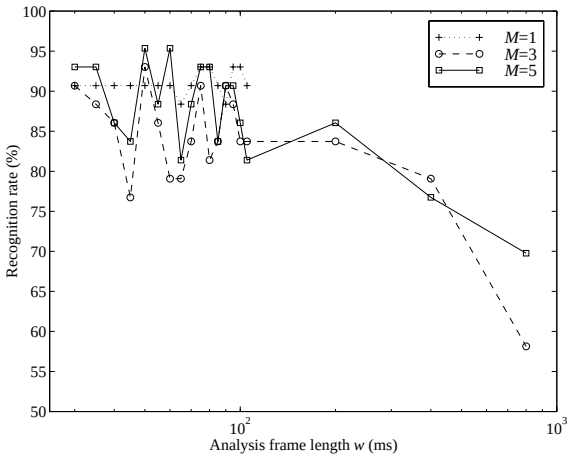

Figure 1: Effect of the analysis frame length w on the recognition rate for $M = 1, 3, 5$

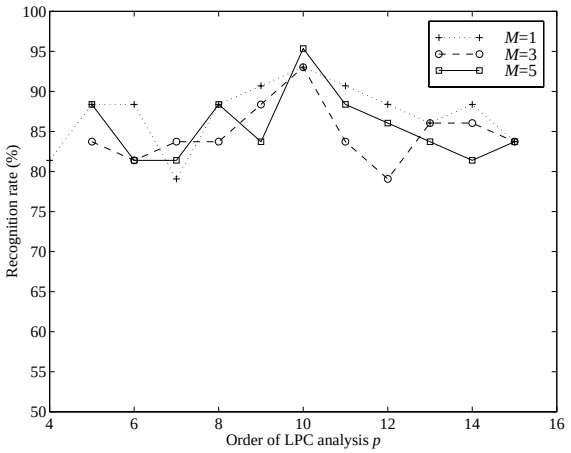

Figure 2: Effect of the order of the autoregressive pre-processor p on the recognition rate for $M = 1, 3, 5$

HMM's. All the other parameters are the same as in the first experiment. It is observed that the best classification results are obtained for frame lengths w larger than the standard speech recognition frame length of 30 ms: on the order of 50–60 ms for three or five-state HMM's, and 80–100 ms for single-state HMM's. This can be interpreted as an indication that the "typical duration" of the acoustic events and transitions is longer in noise events than in speech.

Figure 2 shows the influence of the order p of the LPC analysis on the recognition rate for single-state, three-state, and five-state HMM's. The frame length w was set to 100 ms for the single-state HMM's, to 50 ms for the three-state HMM's, and to 60 ms for the five-state HMM's, respectively. Again, all the other parameters are the same as in the first experiment. The best results are always obtained for $p = 10$, the same value as in speech recognition. The codebook size L was also varied. Classification results for various combinations codebook sizes L, number of states M, and analysis frame length w (in ms) are given in table 1. The best results are always obtained for $L = 256$, for all analysis frame lengths and for

Table 1: Effect of the codebook size L on the number of correct classifications (out of 43)

L	$M = 1$ $w = 80$	$M = 1$ $w = 100$	$M = 3$ $w = 50$	$M = 5$ $w = 50$	$M = 5$ $w = 60$
64	36	34	37	35	36
128	38	38	35	37	37
256	40	40	40	41	41
512	37	35	36	38	34

Table 2: Effect of the analysis frame length and number of states on the number of correct classifications (out of 43) for the one-third-octave pre-processor

w (ms)	$M = 1$	$M = 3$	$M = 5$	$M = 5^*$
100	34	38	36	—
250	—	36	—	—
500	37	38	30	36

all number of states tested. Overall, the best performance achieved was 95% (41/43) correct classifications by a five-state HMM with an analysis frame of 50 ms or 60 ms and with a LPC analysis of order 10 used for the pre-processor.

In the final series of experiments, the LPC-cepstral pre-processor was replaced by the standard one-third-octave filter bank that is used in noise monitoring applications. In this way, it was possible to investigate the possibility of using a HMM-based as a "post-processor" for a standard sound level meter. Table 2 summarizes the results that have been obtained for various analysis frame lengths w (integration times for L_{eq} in noise control parlance). The codebook size was again set to 256. The last column ($M = 5^*$) corresponds to a HMM with five states but with only one "silence" state at the beginning and at the end, instead of three. The best performance achieved was 88% (38/43) correct classification for a three-state HMM and an analysis frame of 100 ms or 500 ms.

4.4. Discussion

Even if the limited size of the MADRAS database means that the performance numbers obtained must be taken carefully, several conclusions can still be drawn. First, it appears that HMM-based classifiers outperform simple spectrum-based classifiers. Indeed, HMM-based classifiers yield more than 90% of correct classifications and even more than 95% for the best of them. On the other hand, spectrum-based classifiers achieve only more than 80% of correct classifications for a similar recognition task on the MADRAS database [2, 1].

The performance improvement shown by HMM-based classifiers could have been expected because the HMM-based classifiers take into account the temporal structure of the noise events unlike the previous spectrum-based classifiers.

Second, the analysis frame length is larger in the noise recognition case. This can be interpreted as an indication that the "typical duration" of the acoustic events and transitions is longer in noise events than in speech.

Third, it seems that the filter bank-based classifier is outperformed by the LPC-based classifier. Interestingly, it can be noted

Table 3: Correct classification by human listeners

Sound	Recognition rate (%)
Car	100
Truck	84.6
Moped	95.8
Aircraft	90.4
Train	92.5

this is also usually the case in speech recognition [3]. This can mean that LPC-cepstral analysis is better suited to noise recognition than one-third-octave analysis. However, this could also be due to the fact that the filter bank pre-processor provides 21-dimensional feature vectors before quantization whereas the LPC-cepstral pre-processor provides 12 coefficients. It is thus possible that, because of the limited size of the training data, the codebook might not be as well trained in the filter bank case as it is in the LPC-cepstral case.

Finally, it is possible that the limited size of the MADRAS database may also have caused training and testing problems.

4.5. Listening Tests

In order to get a "baseline" performance level for a human listener, a series of informal listening tests was conducted. In these tests, human subjects were asked to classify noise event recordings into one of five possible categories. The noise event recordings were the same as the one used in the automatic recognition experiments of the previous section.

For our experiments, 110 noise event recordings were extracted from the MADRAS database, with approximately an equal number of "car," "truck," "moped," "aircraft," and "train" events. The event recordings (WAV files) were played in random order on loudspeakers via the sound board of a PC and a power amplifier. A group of six human listeners was asked to perform the classification test. The listeners' group included engineers with and without noise control experience. Globally, the listeners correctly classified 91.8% of the noise events. Table 3 breaks down the results by categories of sound. Additional results can be found in [7].

Comparing the results obtained by the HMM-based classifiers and the results obtained during the listening test, it appears that, globally, the best classifiers outperform the "average" human listener by a few percents. Lookine more closely at the noise events that were missclassified by human listeners and by HMM-based classifiers, it seems that the sounds that create problems to the HMM classifiers are also often the sounds that create problems to the human listeners. So, in a sense, even when committing a classification error, the classifier might still make a "perceptually meaningful" decision.

5. SUMMARY AND CONCLUDING REMARKS

It has been shown how HMM's could be used to build practical noise classifiers based on a time-frequency analysis of the noise signal. The HMM-based approach to noise recognition has been evaluated experimentally for the classification of five types of noise events (car, truck, moped, aircraft, train). The best results obtained were 95.3% of correct classifications for a five-state HMM using LPC-ceptral pre-processsing. For comparison, a classification test has been performed with human listeners for the same data which has shown that the best HMM-based classifier outperformed the "average" human listener who achieves only 91.8% of correct classification for the same task.

Only discrete HMM's have been evaluated for the classification of noise events because they are the simplest type of HMM. It was not possible to evaluate the performance of more complex models such as Gaussian mixtures HMM's or hybrids neural networks/HMM's because there were not enough training data to use these more demanding models. Performance improvment can thus probably be expected once it becomes possible to use these more complex models.

For further research, it would be a good thing to increase the size of the MADRAS database. It should contain more samples of each of the variants of the noise events. Research in environmental noise recognition would greatly benefits from the creation of large size standardized reusable corpora, like the ones used in speech recognition.

6. ACKNOWLEDGMENT

The authors would like to thank Jean Nemerlin of CEDIA for providing access to the MADRAS database.

7. REFERENCES

[1] P. Chapelle, C. Couvreur and L.-M. Croisez, "Experimental Results on Automatic Recognition of Environmental Noise Sources", *ACUSTICA united with acta acustica*, vol. 82, S1, p. S220, Jan. 1996.

[2] D. Dufournet and P. Jouenne, "MADRAS, an intelligent assistant for noise recognition", in *Proc. INTER-NOISE '97*, Budapest, Hungary, Aug. 1997.

[3] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, Englewood Cliff, NJ, 1993.

[4] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Published, Boston/Doordrecht/London, 1992.

[5] L. R. Rabiner, "A tutorial on hidden Markov models and selected application in speech recognition", *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[6] TCTS-Multitel, Faculté Polytechnique de Mons, Mons, Belgium, *Step by Step Guide to Using the Speech Training and Recognition Tool (STRUT)–Users's Guide*, Aug. 1996, [http://tcts.fpms.ac.be/speech/strut.html].

[7] P. Gaunard and C. G. Mubikangiey, "Reconnaissance automatique des bruits environnementaux", undergraduate thesis, Faculté Polytechnique de Mons, Mons, Belgium, June 1996, in French.