A Pattern Recognition System for Environmental Sound Classification based on MFCCs and Neural Networks

F. Beritelli, *Member, IEEE,* R. Grasso
Dipartimento di Informatica e delle Telecomunicazioni
University of Catania - Italy

Abstract—**The paper proposes a study of a background noise classifier based on a pattern recognition approach using a neural network. The signals submitted to the neural network are characterised by means of a set of 12 MFCC (Mel Frequency Cepstral Coefficient) parameters typically present in the front end of a mobile terminal. The performance of the classifier, evaluated in terms of percent misclassification, indicate an accuracy ranging between 73 % and 95 % depending on the duration of the decision window.**

I. INTRODUCTION

Environmental or background noise is a set of sounds generated in an environment and recorded using a microphone. A typical example is the signals recorded by a mobile phone or land line when a person is listening. By means of a Voice Activity Detector (VAD) it is possible to discriminate between voice activity and moment when there is only background noise, and from subsequent analysis it is possible to identify the type of noise present during a conversation. More generally, the use of a sound recognition system can offer concrete potential in several application scenarios: fixed and mobile telephony, speech recognition, forensic speaker identification, acoustic sensors, and surveillance and security applications [1][3][5][7][10][11].
Various solutions have been proposed in the last few years to make speech processing algorithms more robust to background noise [1-9]. Although their performance in the presence of noise has been enhanced considerably, they are still far from offering the same performance levels as are obtained in clean environments. There is still a wide margin for improvement in the robustness to background noise of most speech processing algorithms. The main idea is represented by the block diagram in Fig. 1.
A general speech processing system is based on several analysis stages, each with different functions (e.g. pre-filtering, feature extraction, matching, post-filtering, decision, etc.). According to the characteristics of the background noise it is possible to adapt each or some blocks dynamically, so as to optimize their performance by selecting the best configuration for that particular type of noise. For example, it would be appropriate to adapt the set of parameters typically extracted from the signal, the pre-processing and post-processing filters, the thresholds in the detection blocks, the set of parameters characterizing the matching blocks, etc. Of course, to do so it is essential to know how each block behaves when the type of noise varies, and to have a background noise classification system that can dynamically identify the type of noise involved, or at least the class containing noise of the same kind, i.e. with similar spectral and statistical characteristics. The task of the noise classifier block is therefore to recognize the type of noise allowing the whole system to adapt dynamically.

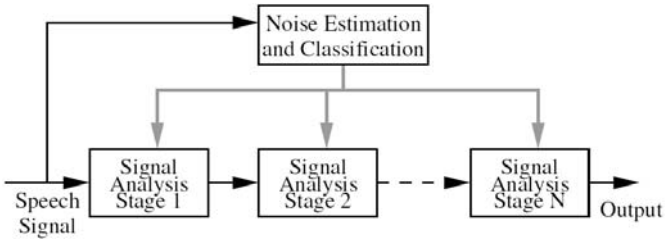

Fig. 1 Adaptive noise robust speech processing system

The approaches analysed up to now are oriented towards the use of pattern recognition techniques based on fuzzy logic [1-2], HMM (Hidden Markov Model) probabilistic models [6][8] and neural networks [5].
In turn, these use different parameters; however, the various parameters proposed in the literature often contrast with currently available technology. This applies in particular to mobile terminals in which, according to the standard currently being used, ETSI ES 202 050, the parameters extracted at the front end are only energy and 13 MFCCs. This paper therefore assesses performance in the recognition of background noise using a subset of these parameters. The aim of the paper is thus to verify the degree of accuracy of a noise classification system based on a subset of 12 MFCC parameters of the ETSI ES 202 050 front-end standard.
Performance was evaluated in terms of the percentage of misclassification using a database of 10 different types of noise in the training and testing phases. The behaviour of the system was also observed when subjected to "external" noise, i.e. when the type of input noise was different from that used in the training phase. The paper is organised as follows. Section 2 presents the architecture of the proposed classifier. Section 3 describes the set of parameters used to characterise the signals, while Section describes the database of background noise used. Sections 5 and 6 give the results of the study and the conclusions drawn.

978-1-4244-4242-3/08/$25.00 ©2008 IEEE

II. Background Noise Classifier

Fig. 2 shows the structure of the classifier presented here. Each signal in the database is first segmented into frames lasting 30 ms, with a partial overlap of 20 ms between consecutive frames. Every 10 ms the set of 12 Mel-Frequency Cepstral Coefficients (MFCCs) are extracted by applying the discrete cosine transform to the log-energy outputs of a Mel-scaling filter-bank. These features are used to describe the spectral shape of a signal. Vector Quantization (VQ) was applied only for training phase.

The matching phase is performed by a Feed-Forward neural network trained by a Resilient Back Propagation algorithm. The network has 12 neurons in the input layer, 24 in the hidden layer and 10 in the output layer.

The final result is entrusted to a logic that decides on the class of noise according to the highest value in the 10 neural network outputs.

III. Parameter Extraction

Each of the signals used underwent pre-processing based on a process of pre-emphasis filtering with a coefficient of $\alpha = -0.95$; fragmentation, in which the signal was subdivided into 30-ms frames with a 20-ms overlap between consecutive frames; and finally the application of a Hamming window to ensure smoothing. 12 MFCC coefficients were extracted from each frame. The choice of the MFCC parameter was linked to the wish to characterise the signals in such a way as to reproduce as closely as possible the perception of noise by the human ear. For the training database, the parameter extraction phase was followed by vector quantisation: this was necessary not only to lighten the computational workload in the supervised training stage but also to reduce the training set.

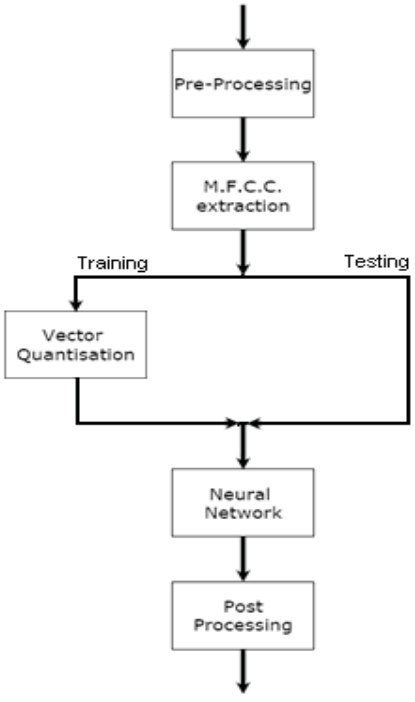

Fig. 2. Architecture of the classifier

8998 frames represented by the 12 coefficients extracted are, in fact, obtained from each signal.

Quantization was performed using a K-means algorithm, with a splitting method and estimation of the Euclidean norm, thus leading to a codebook of 2048 centroids for each class of noise. As can be seen in Tab. 3 the performance of the classifier increases with the size of codebook, but also computational complexity increases with the number of centroids, so a good trade-off was found with 2048 centroids.

IV. Database

The database used in the learning, testing and validation phases contains sequences recorded in various environments, sampled at a frequency of 8 kHz and linearly quantized with 16 bits per sample. More specifically, the training and testing databases comprise 10 recordings lasting 90 seconds each, referring to the 10 classes of noise considered: Bus, Car, Construction, Dump, Factory, Office, Pool, Station, Stadium and Train. The testing sequences are naturally of the same type as the training sequences but refer to different recordings. The database of "external" noise contains 3 classes of noise (Restaurant, Shopping and Street) and each recording lasts 180 seconds (i.e. twice the length of the training and testing sequences).

V. Results

The system underwent two analysis phases, a testing phase based on the use of 10 classes of noise, and a testing phase featuring 3 external noises. Performance was evaluated in terms of accurate recognition of the 10 classes of noise, with a decision window, DW, lasting between 10 ms and 5 seconds.

4 different decision window lengths were taken into consideration:

- 1 frame (10 ms)
- 10 frames (100 ms);
- 100 frames (1 sec.);
- 500 frames (5 sec.)

Analysis of the testing database with a decision window of 10ms (1 frame) shows an average correct classification of 73,24 % (Tab. 1). Particularly high percentages were obtained with the Car (87,6 %), Construction (79,5 %) and Train (95,3 %) classes, whereas for noises identified with less accuracy misclassification tends towards acoustically similar classes. Overall performance was comparable with that obtained in the literature using non-homogeneous sets of parameters, which also require additional implementation and calculation resources as compared with a standard set like MFCC parameters. It can also be seen from Tab. 2 that performance can increase by as much as 20 % in applications where the decision as to the type of noise can be made in a longer time window, which allows a post-processing process to be activated. Finally, Tab. 4 illustrates the behaviour of the noise classifier towards 3 types of noise other than those used during the training phase. The noise frames are assigned to the noise categories present in the classifier and acoustically most similar.

DW=10 ms	Bus	Car	Construct	Dump	Factory	Office	Pool	Station	Stadium	Train
Bus	**68,7%**	2,2%	5,0%	9,2%	0,0%	0,7%	0,5%	0,1%	1,9%	11,6%
Car	2,5%	**87,6%**	0,2%	3,7%	0,0%	0,5%	0,0%	0,0%	0,7%	4,8%
Construction	1,4%	0,0%	**79,5%**	0,1%	3,6%	0,3%	14,2%	0,7%	0,3%	0,0%
Dump	11,4%	3,5%	1,3%	**74,8%**	0,2%	2,2%	0,1%	0,4%	2,0%	4,1%
Factory	0,0%	0,0%	1,6%	0,0%	**71,1%**	8,1%	2,3%	14,7%	2,2%	0,0%
Office	0,7%	0,3%	1,3%	0,2%	17,9%	**58,9%**	2,2%	10,6%	7,6%	0,4%
Pool	0,1%	0,0%	3,9%	0,0%	3,4%	0,6%	**75,3%**	13,1%	3,6%	0,0%
Station	0,0%	0,0%	0,7%	0,0%	9,7%	7,4%	7,1%	**71,4%**	3,7%	0,0%
Stadium	4,6%	0,8%	5,9%	3,5%	3,0%	8,6%	13,2%	8,6%	**49,9%**	1,9%
Train	1,4%	1,9%	0,0%	0,4%	0,0%	0,0%	0,0%	0,0%	1,0%	**95,3%**

Tab. 1: Misclassification matrix: testing sequences

Decision Window	**1 (10 ms)**	**10 (100 ms)**	**100 (1 sec)**	**500 (5 sec)**
Average Accuracy	73,24%	85,06%	92,12%	95,33%

Tab. 2 Average accuracy with varying decision window size

Decision Window	**1 (10 ms)**	**10 (100 ms)**	**100 (1 sec)**	**500 (5 sec)**	**Codebook Size**
Average Accuracy	71,48%	84,21%	91,24%	95,35%	**1024**
	73,24%	85,06%	92,12%	95,33%	**2048**
	75,15%	85,85%	93,25%	96,79%	**4096**

Tab. 3 Average accuracy with varying decision window size and codebook size

DW=10 MS	Bus	Car	Construct	Dump	Factory	Office	Pool	Station	Stadium	Train
Restaurant	3,7%	0,0%	14,2%	0,5%	13,2%	30,5%	8,3%	9,5%	20,0%	0,1%
Shopping	0,1%	0,0%	0,5%	0,1%	10,8%	27,4%	2,3%	51,6%	7,2%	0,0%
Street	18,1%	0,0%	38,0%	1,6%	3,2%	1,7%	19,7%	2,9%	14,8%	0,0%

Tab. 4 Classification of external noises

With reference to the classification of external noise, what was said above was confirmed by a comparison between the power spectrum densities for the categories not used in the training phase (Restaurant, Shopping and Street) and those of the noise classes to which the system assigned higher percentages of misclassification (Office for Restaurant, Station for Shopping and Construction for Street). More specifically, an average power spectrum density was determined for each class of noise (both external and those used to train the system), obtained from the set of power spectrum densities for each frame assigned by the system to a specific class (e.g. all the frames belonging to the Restaurant class classified as Office). Figs. 3 and 4 give an example of the average power spectrum densities obtained from 10 signal frames. Comparison between the "external" and training classes showed that misclassification was dictated by similarities in the signal spectra (Figs. 5 and 6) and thus acoustic similarities.

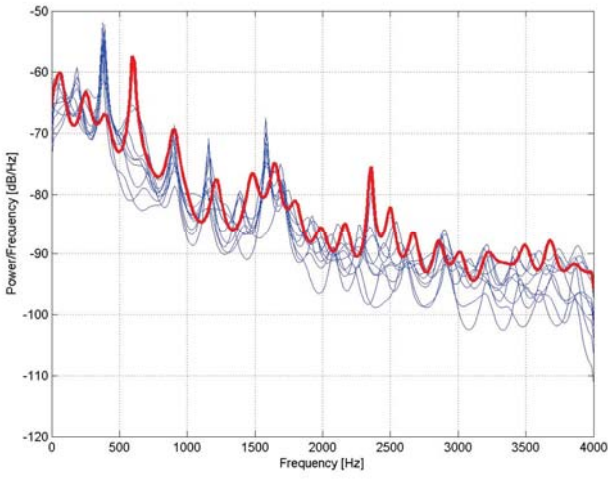

Fig. 3 Average power spectrum densities (restaurant noise)

It was observed that the difference in power within the bands for the extraction of the MFCC parameters is very limited. This means that in the presence of an overlapping speech signal with a fixed average power, in the sub-bands for extraction of the MFCC parameters, the SNR is quite similar in both cases, thus confirming an alteration of the similar SNR by the pairs of classes compared.

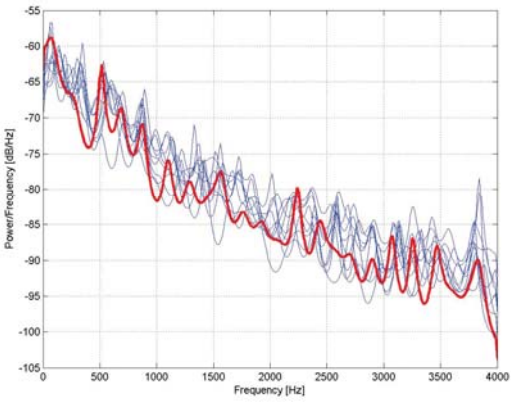

Fig. 4 Average power spectrum densities (shopping noise)

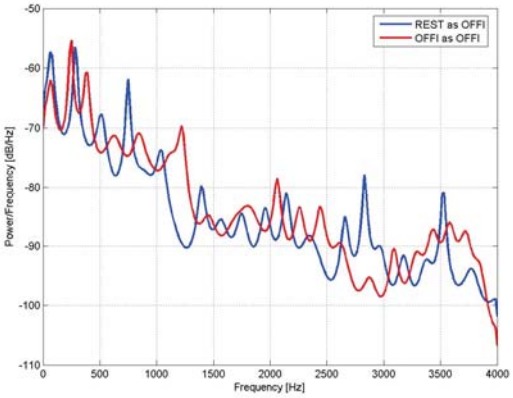

Fig. 5 Average power spectrum densities (office noise – red line - and restaurant noise detected as office noise – blue line-)

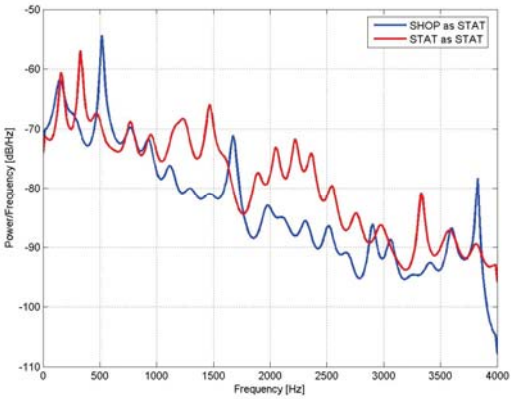

Fig. 6 Average power spectrum densities (station noise – red line - and shopping noise detected as station noise – blue line-)

VI. Conclusions

The paper proposes a background noise classifier using a pattern matching technique based on MFCC parameters and a neural network. The aim was to verify the possibility of achieving performance similar to that of traditional approaches but without requiring the implementation of a specific set and therefore additional calculation resources in a system implementing MFCC parameters by default. The performance obtained, which exhibited an average accuracy ranging between 75 % and about 95 %, is adequate for all the application contexts that allow a response time around 1 second. The response to external noise also indicates that the system is capable of associating an input frame with the noise class that is acoustically most similar.

References

[1] Beritelli F. , Casale S. , “Background Noise Classification in Advanced VBR Speech Coding for Wireless Communications”, IEEE ISPACS, Melbourne, Australia, 1998.

[2] Beritelli F., Casale S., Ruggeri G., “New Results in Fuzzy Pattern Classification of Background Noise”, IEEE ICSP , Beijing, China, 2000.

[3] Betkowska A., Shinoda K., Furui S., “Model Optimization for Noise Discrimination in Home Environment”, Symposium On Large Scale Knowledge Resources, Tokyo, Japan, 2005.

[4] Rabaoui A. , Lachiri Z. , Ellouze N. , "Automatic Environmental Noise Recognition", IEEE ICIT, Tunisia, 2004.

[5] Couvreur L., Laniray M., “Automatic Noise Recognition in Urban Environments Based on Artificial Neural Networks and Hidden Markov Models”, INTERNOISE 2004, Prague, Czech Republic, 2004.

[6] Sabri M., Alirezaie J., Krishnan S., “Audio Noise Detection Using Hidden Markov Model”, IEEE Workshop on Statistical Signal Processing, 28 Sept.-1 Oct. 2003.

[7] Ma L., Smith D. J., Milner P. B., “Context Awareness using Environmental Classification”, EuroSpeech 2003, Geneva, Switzerland, 2003.

[8] Gaunard P., Mubikangiey C. G., Couvreur C., Fontaine V., “Automatic Classification of Environmental Noise Events by Hidden Markov Model”, IEEE ICASSP, 1998.

[9] El Maleh K., Samouelian A. , Kabal P., “Frame-Level Noise Classification in Mobile Environments”, IEEE ISPACS, Phoenix, Arizona/U. S. A., May 1999.

[10] F. Beritelli, S. Casale, S. Serrano, “A Low-Complexity Speech-Pause Detection Algorithm for Communication in Noisy Environments” European Transactions on Telecommunications, Volume 15, Issue 1, January/February 2004, Pages: 33-38

[11] Beritelli, F.; Casale, S.; Serrano, S.; “Adaptive V/UV Speech Detection based on Acoustic Noise Estimation and Classification”, Electronics Letters Volume 43, Issue 4, February 15, 2007, P age(s): 249 – 251.