Environmental Sounds Recognition System, Using the Speech Recognition System Techniques

Omar Aranda , Héctor M. Pérez , Mariko Nakano

Instituto Politécnico Nacional
SEPI-ESIME Culhuacan
Av. Santa Ana, 1000, Col. San Francisco Culhuacan, C. P. 04430, Coyoacan, México D. F.
arandauribe@calmecac.esimecu.ipn.mx

Abstract. This paper describes an environmental sounds recognition system using LPC-Cepstral coefficients as feature vectors and an artificial neural network backpropagation as recognition method. LPC-Cepstral data are totally dependents of the sound-source from which are computed. This system is evaluated using a database containing files from four different sound-sources under a variety of recording conditions. The training patterns used in the network-training ad testing processes, are extracted from the Discrete Fourier transform magnitude of the LPC-Cepstral matrices. The global percentages of verification and identification obtained in the network-testing process are 90.42% and 89.5%. Basically the idea here is to apply the techniques found in speech recognition systems to an environmental sounds recognition system.

Keywords- Artificial Neural Network, LPC-Cepstral Analysis, Discrete Fourier Transform.

Resumen. Este artículo describe un sistema de reconocimiento de sonidos ambientales utilizando como vectores característicos los coeficientes LPC-Cepstral y una red neuronal artificial backpropagation como método de reconocimiento. Los datos LPC-Cepstral son totalmente dependientes de la fuente de sonido de la cual son extraídos. Este sistema es evaluado con una base de datos que contiene archivos de cuatro fuentes de sonido diferentes grabados bajo diversas condiciones. Los patrones de entrenamiento son extraídos de la magnitud de la transformada de Discreta de Fourier. Los porcentajes de verificación e identificación obtenidos en la etapa de prueba de la red son 90.42% y 89.5% respectivamente. Básicamente la idea es aplicar las técnicas utilizadas en los sistemas de reconocimiento de hablante a un sistema de reconocimiento de sonidos ambientales.

Palabras clave- Red Neuronal Artificial, Análisis LPC-Cepstral, Transformada Discreta de Fourier.

1. Introduction.

Signals that humans can hear are one of the most important sources of informatic. Humans obtain much information from not only voices but also non-verbal sounds. panoply of sounds in our daily lives, called "environmental sounds", are important to understand the surroundings. However, they have been little studied except as noi interfering with speech recognition systems. Much less effort has been directed towa systems capable of detecting, isolating, and identifying the panoply of sounds that fill every-day acoustic environment. In recent years, as the development of robots which behave in the real world, machine tools which have the intelligence to look themselves and their peripheral devices, and failure detection in electro domestic devis several studies on recognition of environmental sounds appeared [1-4]. These stud mainly focused on recognition of sound sources.

For recognition, environmental sounds have the following problems to be solved:

1. Environmental sounds are so various and changeful that they are hard to previously.
2. The environmental sounds are not regular in time.

Problem 1 means that we can use the parametric models as strategy for the environmen sounds recognition process, one kind of this parametric model is the artificial neu network backpropagation that uses a supervised learning algorithm. Problem 2 means as in speech recognition systems, sounds must be made regular or stationary on a interval, after this specific features can be extracted from these sounds and a neur network can be trained (recognition process). The idea here is to apply the methodolo found in speech recognition systems to verification and identification of environmen sounds using LPC-Cepstral analysis and an artificial neural network back-propagation recognizing method.

2. Proposed System.

Figure 1 shows the proposed system. This system consist of four sequential processe first a common database of environmental sounds is obtained, after this a segmentatio algorithm is applied to each token (file) of this database; third LPC-Cepstral features extracted from each segmented file and the DFT is computed from these coefficien finally the DFT magnitude is computed and a training strategy is adopted. The decision taken at the final process and a recognition percentage is computed.

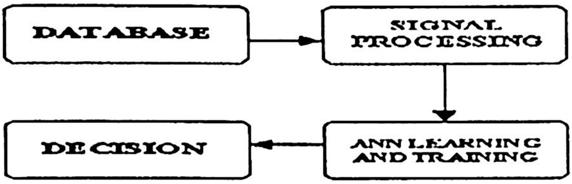

Fig. 1 Proposed environmental sounds recognition system.

2.1 Database Acquisition.

A database containing four different sound-sources was created, files were obtained from an online database. Files were used in the environmental sounds recognition system develop and evaluation. This Database contains 320 files (items). An endpoint algorithm was applied to each signal; this means that we separate portions of the signal stream containing the sound from the portions containing only background noise, which represents computational load to the system. Files are digitalized at 64,000 bits/second. Background sound levels were typically 25 to 30 dB below signal levels.

2.2 Signal Processing.

Figure 2 shows the applied processes in the signal analysis. With this signal analysis a high efficiency of feature extraction is obtained, this facilitates to the neural network the recognition process, this means that higher percentages of verification and identification can be obtained.

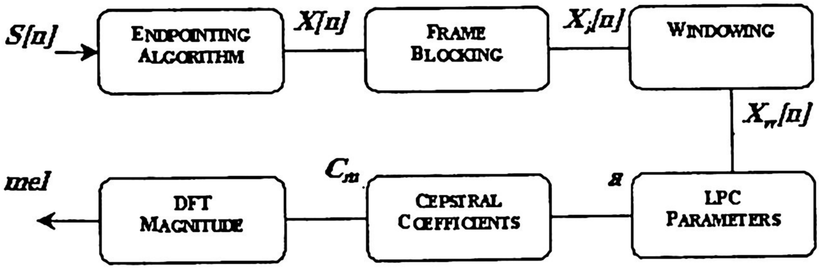

Fig. 2 Shows the methodology used in the signal analysis process

2.2.1 End-pointing Algorithm.

In time domain, magnitude, energy, power, maximums and minimums can be compu from which, the energy is used. Once the energy was calculated, a reference is obtain with this reference the signal can be limited.

In the discrete case the energy is defined as:

$$E[n] = \sum_{n=-\infty}^{\infty} s^2[n] \tag{1}$$

Now, a gamma constant is defined, this constant indicates the number of samples tak from the signal.

In our case the sampling frequency is 8000 Hz. The following step is to make relationship between the sound signal and the gamma constant:

$$E[n] = [(1-\gamma) * E_{n-1}] + [\gamma * y_n^2] \tag{3}$$

To each file stored in the database an end-pointing algorithm was applied. In order to the signal two thresholds of 20 and 10 the maximum energy must be defined, corresponds to the percentage taken form the signal. This algorithm compares thresholds with each sample of the energy until a sample is greater or equal to the thresholds, indicating the signal's beginning.

2.2.2 Frame Blocking and Windowing.

The sound signal, $\hat{S}[n]$, is blocked into frames of 240 samples that corresponds to msec, in which voice is considered stationary [6], with adjacent frames being separate by 120 samples. The use of frames implies three parameters: frame size, frame increme and frame overlapping:

$$S_f = I_f + O_f \tag{4}$$

To the sequence of analysis frames generated from each end-pointed file was applied windowing algorithm, this means that a 240-point Hamming window was used:

$$\hat{S}_w[n] = \hat{S}[n]\,\mathrm{W[n]} \tag{5}$$

Where $0 < n < N - 1$, N is the number of samples in the analysis frame (240 samples) and $W[n]$ is a Hamming window. The frame advancement rate was chosen to yield frames that overlapped at least 50%, and so that the total number of frames between the signal endpoints was at least 64, specifically one second of each signal was analyzed. We used a Hamming window, a typical window used for the autocorrelation method of LPC. This windowing has repercussion in the time responses of the algorithms used but the recognition percentages are improved [6].

2.2.3 LPC Parameters and LPC-Cepstral Coefficients.

In each window 17 LPC coefficients were calculated with Levinson-Durbin recursion. LPC-Cepstral coefficients can be derived directly from the set of LPC coefficients using the recursion:

$$C[n] = -a[n] - \frac{1}{n}\sum_{k=1}^{n=1} kC[k]a[n-k] \tag{6}$$

Where $n > 0$, $C_0 = a_0 = 1, k > p$ and $a[n]$ represents the linear prediction coefficients. The number of frames generated for each signal was of 64. The result in effect was that each signal was represented by a 17 by 64 array of Cepstral coefficients, with the 64 rows representing time and the 17 columns representing frequency.

A 64-point DFT was then calculated for each column 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 column corresponds to a particular spectral frequency, and ach row corresponds to a temporal frequency. The first column contains the DFT of the power envelope of the signal. The first row contains the DFT of the average signal spectrum. The first element of the first column contains the average signal power level. It's typical of two-dimensional Cepstral representations of acoustic signals, and certainly for our signals, that this corner element is the largest component and the size of the components in the first row and first column are larger than the size of interior matrix components. After this the DFT magnitude for each column in the matrix is computed. The LPC-Cepstral coefficients, which are the Fourier transform representation of the spectrum, have been shown to be more robust for speech recognition than the LPC coefficients; in this case we applied this method.

2.3 Feature Extraction and Neural Network Learning and Training.

Sets of two coefficients were selected from the 1024 elements of each DFT magnitu matrix to serve as feature vectors for use in sounds verification and identification. coefficients chosen were take form the first and second columns of the two dimensio Cepstral matrices.

The used model is an artificial neural network backpropagation. The traditional backpròpagation algorithm [5] is used. For each sound pattern, 50 sound files were in the network training process. The sound samples are first normalized so that average magnitude becomes zero and the standard deviation is one. Clusters, or class were formed by grouping the feature vectors for each type of sound. For the netwo training, the ideal number of hidden-layer neurons was chosen from the experimen work. The hope, of course, is that all the samples of each sound will cluster together that space and that cluster for different sounds will be rejected.

3. Results.

Two neural networks, per sound-source, were trained, because of we used two featur vectors from the DFT magnitude matrix of each sound-source. Four stages (one per neur network) were necessary for the network training and each stage corresponds to each stored in the database. 32 input-layer neurons were necessary for the neural netwo training, 10, 15 and 20 hidden-layer neurons were used in this neural network and the results were obtained with 20 neurons; 1 output-layer neuron, per network, was necessa for verify the source-sounds and 4 output-layer neurons in the identification process. training process for verification and identification consist of a matrix with the traini patterns, each network had to be trained with all patterns from all sound-sources.

3.1 Neural Network Testing.

Once the network has been trained, is used in the verification and identification process in this case, the sounds produced by cars, boats, motorcycles and airplanes. verification and identification percentages for each tested neural network can visualized in table 1 and table 2. Sounds-sources that have similarity, as motorcycles, car and airplanes, present higher percentages of false verification than those that don't hav similarities.

Table 1. Percentage of Verification for each Artificial Neural Network (ANN).

Sound-Sources	Percentage of Verification Corresponding to Testing Patterns					
	Training Patterns	Partial Total %	ANN Boats	ANN Airplanes	ANN Cars	ANN Motorcycles
Boat's Sounds	100%	98.33%	96.66%	0%	0%	0%
Airplane's Soynds	100%	88.33%	3.33%	76.66%	10%	10%
Car's Sounds	100%	86.7%	3.33%	10%	73.4%	13.33%
Motorcycle's Sounds	100%	88.33%	0%	6.66%	10%	76.66.%
% False Verification			2.22%	5.55%	6.66%	7.77%
Global % of verification.		90.42%				

The percentage of verification that corresponds to the neural network of boats (96.66%), 4.44% corresponds to patterns verified as false, this means that the output-layer neuron is zero.

Table 2. Percentages of Identification for each sound source.

Sound Sources	Testing Patterns	% identification Training Patterns	% identification Testing Patterns	Total
Airplane's Sound	30	100%	78%	89%
Motorcycle's Sound	30	100%	76%	88%
Car's Sound	30	100%	79%	89.5%
Boat's Sound	30	100%	83%	91.5%
Global %		100%	79%	89.5%

Some LPC-Cepstral coefficients are illustrated in Fig. 3 to 6 and here is demonstrated the difference between sound-sources and the similarity between sounds that come from the same source. Those differences facilitate to the neural network the verification and identification processes.

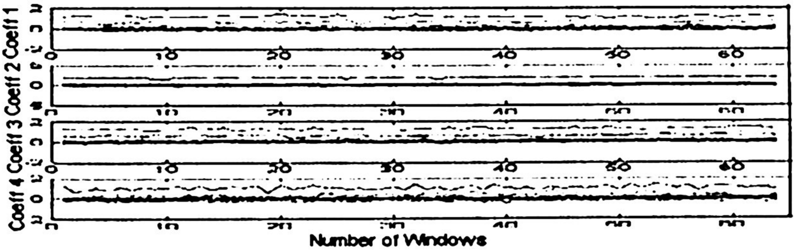

Fig. 3 LPC-Cepstral Coefficients from airplane engine.

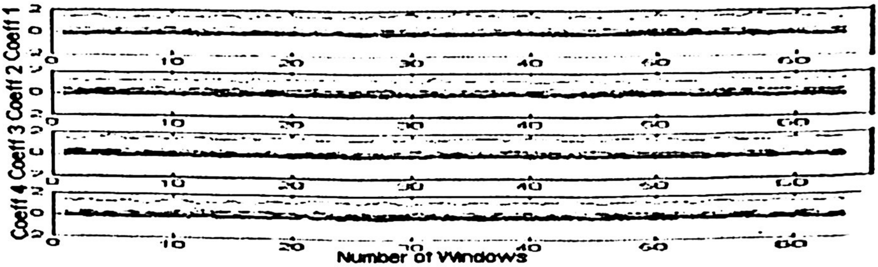

Fig. 4 LPC-Cepstral Coefficients from boats impeller.

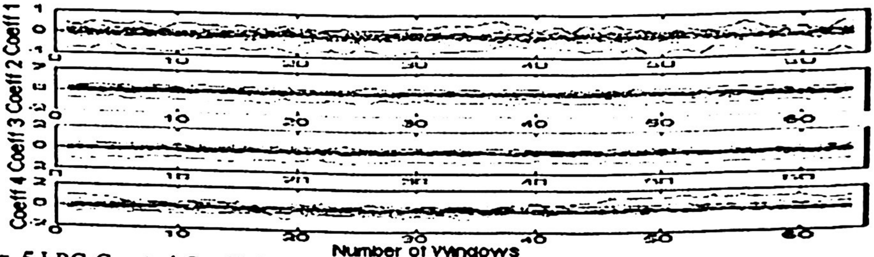

Fig. 5 LPC-Cepstral Coefficients from car engine.

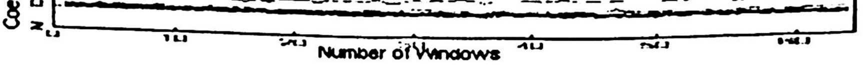

Fig. 6 LPC-Cepstral Coefficients from motorcycle engine.

4. Conclusions.

In this paper was proposed an environmental sounds recognition system based in the LPC Cepstral coefficients feature extraction, after this was computed the DFT magnitude this coefficients matrix. With this matrix an artificial neural network backpropagation trained. The verification and identification percentage were acceptable, 90.42%

89.5% although the number of feature vectors was small; specifically two feature vectors were used. The lowest percentages were obtained for seemed sound-sources, as cars, motorcycles and airplanes. This system seems to be good for some practices applications.

References.

[1] Goldhor, R. S., "Recognition of Environmental Sounds", Proceedings of ICASSP, Vol. 1, pp.149-152, 1993.

[2] Martin, K., "Sound-source recognition: Atheory and computational model", Ph.D. Tesis. *MIT Media Lab, 1999.*

[3] Yuya Hattori, Kazushi. Ishihara, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno, "Repeat Recognition for environmental sounds", in proc. of IEEE International Workshop on Robot and Human Interaction (ROMAN 2004), pp. 83-88, Kurashiki, Sep. 2004.

[4] Yasuhiro Ota, Bogdan M. Wilamoski, "Identifying Cutting Sound Characteristics in Machine Tool Industry with a Neural Network",

[5] Haykin Simon, "Neural Networks A Comprehensive Foundation", edited by Marcia Horton Bayani Mendoza de Leon, Prentice Hall,1999.

[6] 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, 1998.