ISCA Archive
http://www.isca-speech.org/archive

SPECOM'2004: 9th Conference
Speech and Computer
St. Petersburg, Russia
September 20-22, 2004

A System for the Processing of Infant Cry to Recognize Pathologies in Recently Born Babies with Neural Networks

Orion F. Reyes-Galaviz & Carlos Alberto Reyes-Garcia***

*Instituto Tecnologico de Apizaco, Av. Tecnologico S/N, Apizaco, Tlaxcala, 90400, Mexico
orionfrg@yahoo.com
**Instituto Nacional de Astrofisica Optica y Electronica, Luis E. Erro 1, Tonantzintla, Puebla, 72840, Mexico
kargaxxi@inaoep.mx

Abstract

In this work we present the design of an automatic infant cry recognition system that classifies three different kinds of cries, which come from normal, deaf and asphyxiating infants, of ages from one day up to nine months old. The classification is done through a pattern classifier, where the crying waves are taken as the input patterns. We have experimented with patterns formed by vectors of Mel Frequency Cepstral Coefficients and Linear Prediction Coefficients. The acoustic feature vectors are then processed, to be classified in their corresponding type of cry, through an Input Delay Neural Network, trained by gradient descent with adaptive learning rate back propagation algorithm. To perform the experiments and to test the recognition system, we train the neural network with cries from randomly selected babies, and test it with a separate set of cries from babies selected only for testing. Here, we present the design and implementation of the complete system, as well as the results from some experiments, which in the presented case are up to 86 %.

1. Introduction

It has been found that the infant's cry has much information on its sound wave. For small infants, crying is a form of communication, a very limited one, but similar to the way an adult communicates. Given that the crying in babies is a primary communication function, governed directly by the brain, any alteration on the normal functioning of the babies' body is reflected in the cry. The pathological diseases in infants are commonly detected several months or years, after the infant is born. If any of these diseases could be detected earlier, they can be attended and maybe avoided by the opportune application of treatments and therapies [1]. Based on the information carried by the crying wave, the infant's physical state can be determined, and the physical pathologies detected, since very early growing stages. Given that the processing of the information in the infant cry is basically a kind of pattern recognition, we approached the task by using the same techniques used for automatic speech recognition. This work presents the design and implementation of a system that classifies three different kinds of cries. They are recordings of normal, deaf and asphyxiating infants, of ages from 1 day up to 6 months old. For the acoustic processing we used Praat [2]. Additionally, to classify the infant's cry information, we used Matlab to build an input delay neural network. In the model here presented, we classify the original input vectors, without reduction, in three corresponding classes, normal cry, hypoacoustic (deaf) and asphyxiating cries. We train the neural network using a randomly selected set of infant cry samples, from a pre established set of babies, and we test the system using infant cry samples from a completely different set of babies. Here we show scores that go from 71 % up to 86.06 % in precision on the classification. On the next sections we make a description of the full automatic recognition process, followed by the methods we used to perform the extraction of the acoustic characteristics, along with definitions of the proposed system, and the used algorithms. On section three we show the developed system as well as the results obtained. And finally, we present the conclusions and main references.

2. Importance of Early Clinical Diagnosis of Deafness and Asphyxia

The area of neuro-linguistics considers the just born baby cry as the first linguistic manifestation. It is the first experience in the sound production, which is followed by propioceptions of laryngeal and oral movements, altogether with the listening capabilities feedback. This, very soon, will be used to the production of phonemes. Late diagnosis of deafness is proven to cause delay in language development. Children with hearing losses identified before 6 months of age have significantly better language development than children whose hearing losses are identified after 6 months of age. Children with normal cognitive development whose hearing losses are identified before six months can develop language at the same or a similar rate to a hearing child [3]. Children identified with a hearing loss between birth and six months old have a receptive language of 200 words and expressive language of 117 words, whereas those identified between ages of seven and 18 months have a receptive language of 86 words and expressive language of 54 words. When tested at 26 months of age, those identified as deaf before six months old have "significantly higher" measures of language growth and personal-social development [4].

In recent studies the subjective auditory analysis of voice and speech, have been replaced by objective spectral-phonographic analysis, with the help of audio taping and computerized analysis. Special attention has been placed in the spectral-phonographical analysis of the just born baby crying. Preliminary reports are focused in the analysis of full term newborns, premature, with neurological, metabolic, or chromosomal alterations just born babies, as well as those with congenital anomalies among others. The studies suggest an objective variability of the crying related to the health state and the neuropsychological integrity. In the case of the infants who have been through a period of new-born asphyxia, they are exposed to possible changes or neurological level disturbance, depending on the degree of asphyxia that they had suffered. According to the American Academy of Pediatrics (AAP), about 2 to 6 out of 1000 full term newborns present asphyxia, and the incidence is of 60 % in premature newborns with low weight. From them, about 20 to 50 % die during the neonatal period. From the survivors, 25 % develop permanent neurological sequels.

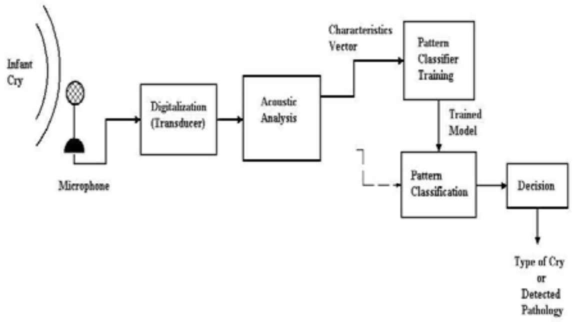

Figure 1. Automatic Infant Cry Recognition Process

3 The Infant Cry Automatic Recognition Process

The infant cry automatic classification process is, in general, a pattern recognition problem, similar to Automatic Speech Recognition (ASR). The goal is to take the wave from the infant's cry as the input pattern, and at the end obtain the kind of cry or pathology detected on the baby [5], [6]. Generally, the process of Automatic Cry Recognition is done in two steps. The first step is known as signal processing, or feature extraction, whereas the second is known as pattern classification. In the acoustical analysis phase, the cry signal is first normalized and cleaned, and then it is analyzed to extract the most important characteristics in function of time. Some of the more used techniques for the processing of the signals are those to extract: linear prediction coefficients, cepstral coefficients, pitch, intensity, spectral analysis, and Mel's filter bank. The set of obtained characteristics is represented by a vector, which, for the process purposes, represents a pattern. The set of all vectors is then used to train the classifier. Later on, a set of unknown feature vectors is compared with the knowledge that the computer has to measure the classification output efficiency. Figure 1 shows the different stages of the described recognition process.

4 Acoustical Processing

The acoustical analysis is the process through which the acoustical features are extracted from the crying wave, the process also implies the application of normalization and filtering techniques, segmentation of the signal, and data compression. With the application of the selected techniques the goal is to describe the signal in terms of some of its fundamental components. A cry signal is complex and codifies more information than the one needed to be analyzed and processed in real time applications. For that reason, in our cry recognition system we used an extraction function as a first plane processor. Its input is a cry signal, and its output is a vector of features that characterizes the key elements of the cry's sound wave. The set of vectors obtained this way are later fed to a recognition model, first to train it, and later to classify the type of cry. After experimenting with diverse types of acoustic characteristics, we report in this work the results obtained with the Mel Frequency Cepstral Coefficients and Linear Prediction Coefficients.

4.1 MFCC (Mel Frequency Cepstral Coefficients)

The Mel spectrum operates on the basis of selective weighing of the frequencies in the power spectrum. High order frequencies are weighed on a logarithmic scale whereas lower order frequencies are weighed on a linear scale. The Mel scale filter bank is a series of L triangular bandpass filters that have been designed to simulate the bandpass filtering believed to occur in the auditory system. This corresponds to series of bandpass filters with constant bandwidth and spacing on a Mel frequency scale. On a linear frequency scale, this spacing is approximately linear up to 1KHz and logarithmic at higher frequencies (see Figure 2). Most of the recognition systems are based on the MFCC technique and its first and second order derivative. The derivatives normally approximate trough an adjustment in the line of linear regression towards an adjustable size segment of consecutive information frames. The resolution of time and the smoothness of the estimated derivative depend on the size of the segment [7].

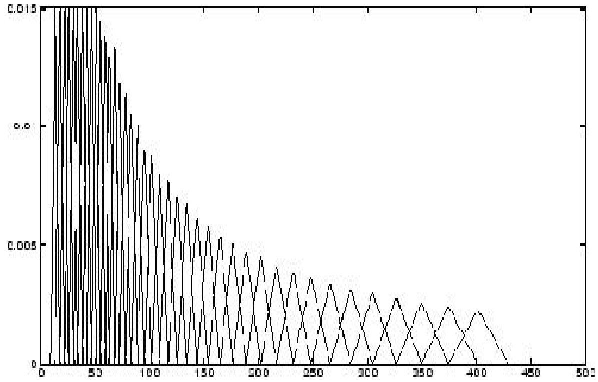

Figure 2. The MFCC filter bank.

4.2 LPC (Linear Prediction Coefficients)

Linear Predictive Coding (LPC) is one of the most powerful speech analysis techniques, and one of the most used methods to encode good quality speech at a low bit rate. It provides extremely accurate estimates of speech parameters, and is relatively efficient for computation. Based on these reasons, we are using LPC to represent the crying signals. Linear prediction is a mathematical operation where future values of a digital signal are estimated as a linear function of previous samples. In digital signal processing linear prediction is often called linear predictive coding (LPC) and can thus be viewed as a subset of filter theory. In system analysis (a subfield of mathematics), linear prediction can be viewed as a part of mathematical modeling or optimization [8].

5. Cry Patterns Classification

The set of acoustic characteristics obtained in the extraction stage, is represented generally as a vector, and each vector can be taken as a pattern. These vectors are later are used to make the classification process. There are four basic schools for the solution of the pattern classification problem, those are: a) Pattern comparison (dynamic programming), b) Statistic Models (Hidden Markov Models HMM). c) Knowledge based systems (expert systems) and d) Connectionists Models (neural networks) [9]. For the present work development we'll focus on the description of the connectionist models, known as neural networks. We have selected this kind of models, in principle, because of their adaptation, simplicity and learning capacity. Besides, one of their main functions is pattern recognition, this kind of models are still under constant experimentation, but their results have been very satisfactory.

5.1 Neural Networks

The neural networks have been defined as systems composed of many simple processing elements, that operate in parallel and whose function is determined by the network's structure, the strength of its connections, and the processing carried out by the processing elements or nodes. Generally, the neural networks are adjusted or trained so that an input in particular leads to a specified or desired output. The training of a network is done trough changes on the weights based on a set of input vectors. The training adjusts the connection's weights from the nodes, after obtaining an output from the network and comparing it with a wished output, with previous presentation of the whole set of input vectors. The neural networks have been trained to make complex functions in many application areas including the pattern recognition, identification, classification, speech, vision, and control systems. In general, the training can be supervised or not supervised. The methods of supervised training are those that are more commonly used, when labeled samples are available. Among the most popular models are the feed-forward neural networks, trained under supervision with the back-propagation algorithm [9]. For the present work we have used a variation of this basic model which is described below.

5.2 Feed Forward Input Delay Neural Network

Cry data are not static, and any cry sample at any instance in time is dependent on crying patterns before and after that instance in time. A common flaw in the traditional Back-Propagation algorithm is that it does not take this into account. Waibel et al. set out to remedy this problem in {10} by proposing a new network architecture called the ``Time-Delay-Neural Network''\ or TDNN. The primary feature of TDNNs is the time-delayed inputs to the nodes. Each time delay is connected to the node via its own weight, and represents input values in past instances in time. TDNNs are also known as Input Delay Neural Networks because the inputs to the neural network are the ones delayed in time. If we delay the input signal by one time unit and let the network receive both the original and the delayed signals, we have a simple time-delay neural network. Of course, we can build a more complicated one by delaying the signal at various lengths. If the input signal is *n* bits and delayed for *m* different lengths, then there should be *nm* input units to encode the total input. When new information arrives, it is placed in nodes at one end and old information shifts down a series of nodes like a shift register controlled by a clock [11]. A general architecture of time-delay networks is drawn in Figure 3.

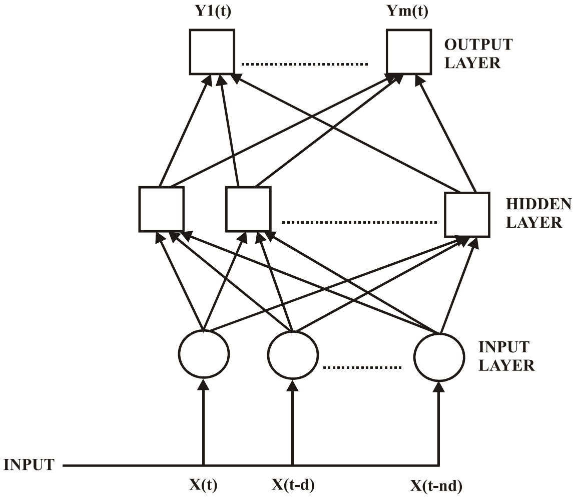

Figure 3. A time delay neural network whose inputs contain a number of tapped delay lines.

The Feed-forward input delay neural network consists of *N1* layers that use the dot product weight update function, which is a function that applies weights to an input to obtain weighed entrances. The first layer has weights that come from the input with the input delay specified by the user, in this case the delay is *[0 1*]. Each subsequent layer has a weight that comes from a previous layer. The last layer is the output of the network. The adaptation is done by means of any training algorithm. The performance is measured according to a specified performance function. Some of the most notorious properties of TDNNs are: i) The Network is shift-invariant: A pattern may be correctly recognized and classified regardless of its temporal location. ii) The network is not sensitive to phoneme boundary misalignment: The TDNN is not only able to learn from badly aligned training data, it is even able to correct the alignment. It does this by learning where the phoneme's presence is significant within the segment of speech. This property is later used to perform recursive sample re-labeling. iii) The network requires small training sets. In [12] Tebelskis quotes the findings of several papers that indicate that the TDNN, when exposed to time-shifted inputs with constraint weights, can learn and generalize well even with limited amounts of training data.

5.2 Training by Gradient Descent With Adaptive Learning Rate Back-propagation

The training by gradient descent with adaptive learning rate back-propagation, proposed for this project, can train any network as long as its weight, net input, and transfer functions have derivative functions. Back-propagation is used to calculate derivatives of performance with respect to the weight and bias variables. Each variable is adjusted according to gradient descent. Several adaptive learning rate algorithms have been proposed to accelerate the training procedure. The following strategies are usually suggested: i) start with a small learning rate and increase it exponentially, if successive epochs reduce the error, or rapidly decrease it, if a significant error increase occurs, ii) start with a small learning rate and increase it, if successive epochs keep gradient direction fairly constant, or rapidly decrease it, if the direction of the gradient varies greatly at each epoch and iii) for each weight an individual learning rate is given, which increases if the successive changes in the weights are in the same direction and decreases otherwise. Note that all the above mentioned strategies employ heuristic parameters in an attempt to enforce the monotone decrease of the learning error and to secure the converge of the training algorithm [13].

6. System Implementation

On the first stage, the infant's cries are collected by recordings obtained directly from doctors of the Instituto Nacional de la Comunicacion Humana (Mexican National Institute of the Human Communication) and IMSS Puebla (Mexican Institute of Social Security). This is done using SONY (ICD-67) digital recorders. The cries are captured and labeled orally at the end of each cry recorded, by the doctor taking the recording, and then labeled in the computer using this label. Later, each signal wave is divided in segments of 1 second, and each one constitutes a sample. From here, for the present experiments we have a corpus made up of 1049 samples of normal infant cry, 879 of hypo acoustics, and 340 with asphyxia. At the following step the samples are processed one by one extracting their acoustic characteristics, LPC and MFCC, by the use of the freeware program Praat 4.0 [2]. The acoustic characteristics are extracted as follows: for every second we extract 16 coefficients from 100-millisecond frames, generating vectors with 144 coefficients by sample. To perform the experiments and to test the recognition system, we wanted to train the neural network with cries from randomly selected babies, and test it with separated cries selected only for testing. In other words, we wanted to test the neural network with cries that the system has never been trained with. In order to do this, we first choose crying samples from some babies for training, and later we choose other ones for testing. We ended up with 285 crying samples from asphyxiating babies for training and 55 for testing; 792 crying samples from normal babies for testing and 257 for testing; and finally 585 crying samples from deaf babies for training and 240 for testing.

The neural network and the training algorithm are implemented with the Matlab's Neural Network

Toolbox. The neural network's architecture consists of 144 neurons on the input layer, a hidden layer with 60 neurons, and one output layer with 3 neurons. The delay used is [0 1]. In order to make the training and recognition test, we select 285 samples randomly on each class. The number of asphyxiating cry samples available determines this number. With these vectors the network is trained. The training is made until 2000 epochs have been completed or an error of 1x 10^{-6} has been reached. After the network is trained, we test it with the 55 samples of each class randomly selected from the crying samples separated for testing, the asphyxiating cry samples will always be the same, this until we have more samples from this pathological cry. The recognition accuracy percentage, from each experiment, is presented in a confusion matrix.

7. Experimental Results

The classification accuracy was calculated by taking the number of samples correctly classified, divided by the total number of samples. The detailed results of two tests of each kind of acoustic characteristic used, LPC and MFCC, with samples of 1 second, with 16 coefficients for every 100 ms frame, are shown in the following confusion matrices.

Results from two experiments using LPC.

	normal	deaf	asphyxia
normal	32	3	11
deaf	0	55	0
asphyxia	15	0	40

Table 1. Confusion matrix showing a 76.06 % precision, after 2000 training epochs and an error of 1x 10^{-2}

	normal	deaf	asphyxia
normal	37	6	9
deaf	0	55	0
asphyxia	30	0	25

Table 2. Confusion matrix showing a 71 % precision, after 2000 training epochs and an error of 1x 10^{-2}

Results from two experiments using MFCC.

	normal	deaf	asphyxia
normal	32	3	11
deaf	0	55	0
asphyxia	15	0	40

Table 3. Confusion matrix showing a 86.06 % precision, after 1414 training epochs and an error of 1x 10^{-6}

	normal	deaf	asphyxia
normal	32	3	11
deaf	0	55	0
asphyxia	15	0	40

Table 4. Confusion matrix showing a 84.24 % precision, after 1443 training epochs and an error of 1x 10^{-6}

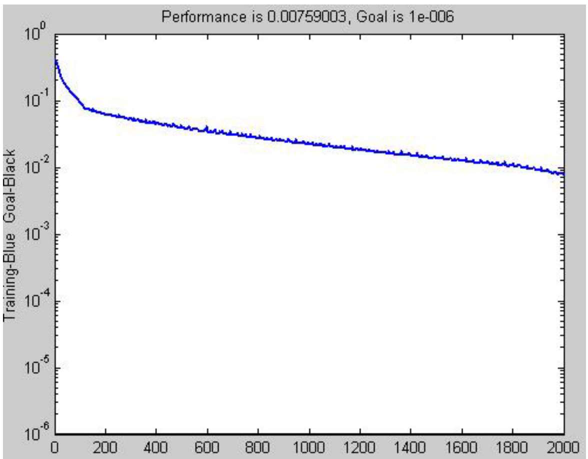

Figure 4. Training with LPC vectors

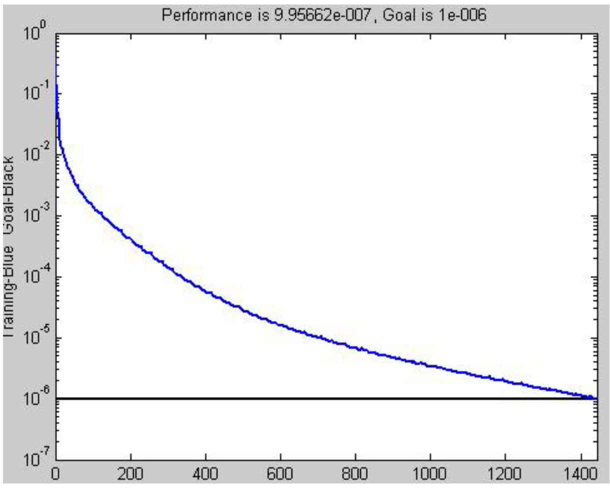

Figure 5. Training with MFCC vectors

7.1. Results Analysis

As can be observed from Figure 4 and Figure 5, the training of the neural network with LPC is slower that the one done with MFCC. With LPC features, the neural network stops until it reaches the 2000 epochs we defined, however the error only goes just over 1x 10^{-2}. On the other hand, with MFCC, the network converges when it reaches the defined error, that is 1x 10^{-6}, and after the training has reached only around 1400 epochs. With the reported experiments we confirm that it is easier to classify the tested cry patterns extracting MFCC features. Until now, we suppose that these results might be due to the fact that MFCC features are easier to differentiate, for the kind of neural network used, than the LPC ones. We are still in the process of analyzing this behavior. We don't have to discard the fact that the training with LPC characteristics gave good results, the inconvenience was that the process of training was slower, the error was higher, and the classification accuracy was lower compared to the results obtained when using the MFCC features. On the other hand, with these experiments we show that the network not only recognizes the type of cry or pathology on the infant, but also recognizes similar characteristics in cries from different babies.

8. Conclusions and Future Work

This work shows the efficiency of the feed forward input (time) delay neural network, particularly when using the Mel Frequency Cepstral Coefficients. It is also shown that the results obtained, of up to 86.06 %, are a little better or similar than the ones obtained in other previous works mentioned. These results can also have to do with the fact that we use different babies to test the network, with the objective of detecting not only the type of cry, but also the baby with the same pathology as the ones used to train the network. In order to compare the obtained performance results, and to reduce the computational cost, we plan to try the system with an input vector reduction algorithm by means of evolutionary computation. This is for the purpose of training the network in a shorter time, without decreasing accuracy. We are also still collecting well-identified samples from the three kinds of cries, in order to assure a more robust training. Among the works in progress of this project, we are in the process of testing new neural networks, and also testing new kinds of hybrid models, combining neural networks with genetic algorithms and fuzzy logic, or other complementary models.

Acknowledgments

This work is part of a project that is being financed by CONACYT-Mexico (37914-A). We want to thank Dr. Edgar M. Garcia-Tamayo and Dr. Emilio Arch-Tirado for their invaluable collaboration when helping us to collect the crying samples, as well as their wise advice.

3. References

[1] O. Wasz-Hockert, J. Lind, V. Vuorenkoski, T. Partanen y E. Valanne, El Llanto en el Lactante y su Significación Diagnóstica, Cientifico-Medica, Barcelona, 1970.

[2] Boersma, P., Weenink, D. Praat v. 4.0.8. A system for doing phonetics by computer. Institute of Phonetic Sciences of the University of Amsterdam. February, 2002.

[3] Yoshinaga-Itano 1998, quoted in 'The High Cost Of Hearing Lost; What Our Publics Need to Know', Donald Radcliffe, The Hearing Journal, May 1998, vol 51 no. 5.

[4] Yoshinaga-Itano, C. & Appusso, M L., The development of deaf and hard of hearing children identified through the high-risk registry, in The American Annals of the deaf, 143, 416-424. 1998.

[5] Sergio D. Cano, Daniel I. Escobedo y Eddy Coello, El Uso de los Mapas Auto-Organizados de Kohonen en la Clasificación de Unidades de Llanto Infantil, Grupo de Procesamiento de Voz, 1er Taller AIRENE, Universidad Catolica del Norte, Chile, 1999, pp 24-29.

[6] Ekkel, T. "Neural Network-Based Classification of Cries from Infants Suffering from Hypoxia-Related CNS Damage", Master Thesis. University of Twente, 2002. The Netherlands.

[7] Gold, B., Morgan, N. (2000), Speech and Audio Signal Processing. Processing and Perception of Speech and Music. John Wiley & Sons, Inc.

[8] Markel, John D., Gray, Augustine H., (1976). Linear prediction of speech. New York: Springer-Verlag.

[9] Lin Chin-Teng, and George Lee, C.S., Neural Fuzzy System: A Neuro-Fuzzy Synergism to Intelligent Systems, Prentice Hall, Upper Saddle River, NJ, 1996.

[10] Waibel A., Hanazawa T., Hinton G., Shikano K., Lang K., Phoneme Recognition Using Time-Delay Neural Networks, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol 37, No 3, March 1989, pp 328 - 339.

[11] LiMin Fu. (1994), Neural Networks in Computer Intelligence. McGraw-Hill International Editions, Computer Science Series.

[12] Tebelskis J., (1995) Speech Recognition Using Neural Networks, PhD Dissertation, Carnegie Mellon University.

[13] V.P. Plagianakos*, M.N. Vrahatis*, G.D. Magoulas**, Non-monotone Methods for Back-propagation Training with Adaptive Learning Rate; University of Patras*, Patras, Greece; University of Athens** Athens, Greece. Technical Report No TR98-04, University of Patras 1998.