Infant Cry Classification to Identify Hypo acoustics and Asphyxia with an Evolutionary-Neural System

ORION F. REYES GALAVIZ*, CARLOS A. REYES GARCÍA**,
& SERGIO D. CANO ORTIZ***
*Universidad Autonoma de Tlaxcala, Calzada Apizaquito S/N,
Apizaco, Tlaxcala, 90400, MEXICO
**Instituto Nacional de Astrofísica Óptica y Electrónica,
Luis E. Erro 1, Tonantzintla, Puebla, 72840, MEXICO
***Universidad de Oriente
Patricio Lumumba S/N, Santiago de Cuba 90500, CUBA

Abstract: - This work presents an infant cry automatic recognizer development, with the objective of classifying three kinds of infant cries, normal, deaf and asphyxia from Mexican and Cuban recently born babies. We use extraction of acoustic features such as LPC (Linear Predictive Coefficients) and MFCC (Mel Frequency Cepstral Coefficients) for the cry's sound waves, and a genetic feature selection system combined with a feed forward input delay neural network, trained by adaptive learning rate back-propagation. We describe the whole process; in which we include the acoustic features extraction, the hybrid system design, implementation, training and testing. We also show the results from some experiments, in which we obtain up to 94.66% precision.

Key-Words: - Feature Selection, Evolutionary Strategies, Classification, Infant Cry, Pattern Recognition, Hybrid System.

1 Introduction

The cry sound produced by an infant is the result of his/her physical and psychological condition and/or from internal/external stimulation. It has been proved that crying caused by pain, hunger, fear, stress, etc. shows different cry patterns. An experimented mother can be able of recognizing the difference between different types of cry, and with this, react adequately to her infant's needs. The experts in neurolinguistics consider the infant cry as the first speech manifestation. It is the first experience on the production of sounds, which is followed by the larynx and oral cavity movements. All of this, combined with the feedback of the hearing capability, will be used for the phonemes production. Children with hearing loss, identified before their first 6 months of life, have a significant improvement in the speech development than those whose hearing loss was identified after their first 6 months of age. In the case of the infants that have passed through an asphyxiating period at birth, they are exposed to changes in a neurological level, depending on the asphyxiating range that they had suffered. According to the American Academy of Pediatrics (AAP), from 2 to 6 out of 1000 recently born babies' present asphyxia and 60% of the babies prematurely born and presenting low weight also suffer an asphyxiating period. From them, 20 to 50% die during their first days of life. From the survivors, 25% develop permanent neurological sequels.

It has been reported in the literature that the infant cry analysis, initially was done mainly by means of visual comparison of spectrograms. This method is very dependant of subjective evaluation and it's not very appropriate for its generalized use with diagnosis purposes. From the beginning of the 20th century, some efforts have been made to investigate the infant cry and discover how to use the information 'hidden' inside the sound. Therefore, it was necessary the development of objective methods to allow us to automatically analyze infant cry in order to make its correct classification. In this work we show an infant cry unit classifier in which we obtain results up to 94.66% precision.

2. State of the Art

Recently, some research efforts had been made that show interesting results, and highlight the importance of exploring this field. In [1], Reyes & Orozco classify samples of deaf and normal babies, obtaining recognition results that go from 79.05% up to 97.43%. Petroni used neural networks to differentiate pain and

no-pain cries [2]. Tako Ekkel tried to classify newborn crying sound into two categories normal and abnormal (hypoxia), and reports a correct classification result of 85% based on a radial basis neural network [3]. Also, using self organized maps methodologies, Cano et al, in [4] report some experiments to classify infant cry units from normal and pathological babies.

3. INFANT'S CRY AUTOMATIC RECOGNITION PROCESS

The automatic infant's cry recognition process (Fig. 1) is basically a problem of pattern recognition, and similar to speech recognition. The goal is to take the baby's cry sound wave as an input, and at the end obtain the kind of cry or pathology detected. Generally, the Pattern Recognition Process is done in two steps; the first step is the acoustic processing, or features extraction, while the second is known as pattern processing or classification. In the proposed system, we have added an extra step between both of them, called feature selection. For our case, in the acoustic analysis, the infant's cry signal is processed to extract relevant features in function of time. The feature set obtained from each cry sample is represented by a vector, and each vector is taken as a pattern. Next, all vectors go to an acoustic features selection module, which will help us; to select the best features for the training process, and at the same time to reduce the input vectors. The selection is done through the use of evolutionary strategies. As for the pattern recognition methods, four main approaches have been traditionally used: pattern comparison, statistical models, knowledge based systems, and connectionist models. We focus in the last one.

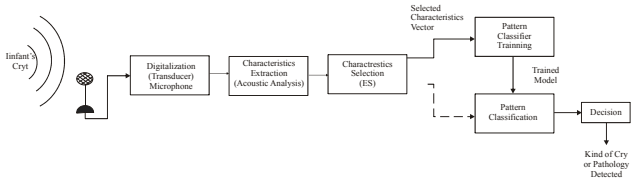

Fig. 1. Infant's Cry Automatic Recognition Process

4. ACOUSTIC PROCESSING

The acoustic analysis implies the application and selection of filter techniques, feature extraction, signal segmentation, and normalization. With the application of these techniques we try to describe the signal in terms of its fundamental components. One cry signal is complex and codifies more information than the one needed to be analyzed and processed in real time applications. For this reason, in our cry recognition system we use a feature extraction function as a first plane processor. Its input is a cry signal, and its output is a vector of features that characterizes key elements of the cry's sound wave. We have been experimenting with diverse types of acoustic features, emphasizing by their utility Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coefficients (LPC).

4.1 Linear Predictive Coefficients

Linear Predictive Coding (LPC) is one of the most powerful techniques used for speech analysis. 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 [1].

4.2 Mel Frequency Cepstral Coefficients

The low order cepstral coefficients are sensitive to overall spectral slope and the high-order cepstral coefficients are susceptible to noise. This property of the speech spectrum is captured by the Mel spectrum. High order frequencies are weighted on a logarithmic scale whereas lower order frequencies are weighted on a linear scale. The Mel scale filter bank is a series of *L* triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass 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 (Fig. 2).

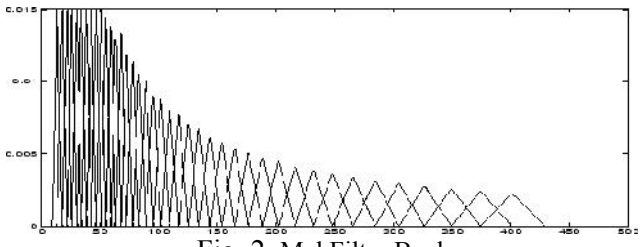

Fig. 2. Mel Filter Bank

Many speech 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 [5].

In conclusion, MFCC can be calculated by:

- Converting the signal in small segments.

- Calculate the DFT for each segment
- The spectrum converts into a logarithmic scale
- The scale is transformed into a soft MEL spectrum
- The discrete cosine transform is calculated (to reduce the spectrum)
- Typically, the first 13 coefficients are used.

5 CRY PATTERN CLASSIFICATION

The acoustic features set obtained in the extraction stage, is generally represented as a vector, and each vector can be taken as a pattern. These vectors are later used to make the acoustic features selection and classification processes. For the present work we are using a classifier corresponding to the type of connectionist models known as neural networks, they are reinforced with evolutionary strategies to select features in order to improve their learning process. Having as a result a *Genetic-Neural* system

5.1 Evolutionary Strategies

The evolutionary strategies are proposed to solve continuous problems in an efficient manner. Its name comes from the German "*Evolutionstrategien*", so we may frequently see them mentioned as "ES". Their origin was an stochastic scaled method (in other words, following the gradient) using adaptive steps, but with time it has converted in one of the most powerful evolutionary algorithms, giving good results in some parametric problems on real domains. The Evolutionary Strategies make more exploratory searches than genetic algorithms [6].

The main reproduction operator in evolutionary strategies is the Gaussian mutation, in which a random value of a Gaussian distribution is added to each element from an individual to create a new descendant (Fig. 3) [7].

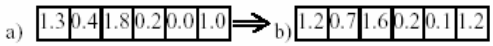

Fig. 3. Gaussial Mutation from parent *a* to produce descendat *b*.

The selection of parents to form descendants is less strict than in genetic algorithms and genetic programming.

5.2 Neural Networks

In a study from DARPA [8] the neural networks are 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. We can train a neural network to execute a function in particular, adjusting the values of the connections (weights) between the elements. Generally, the neural networks are adjusted or trained so that an input in particular leads to a specified or desired output (Fig.4). Here, the network is adjusted based on a comparison between the actual and the desired output, until the network's output matches the desired output [9].

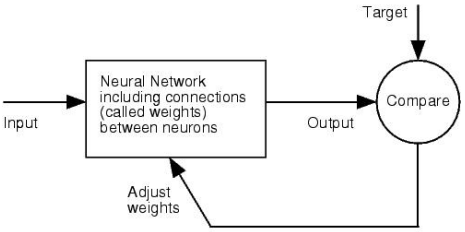

Fig. 4. Tranning of a Neural Network

Generally, the training of a neural network can be supervised or not supervised. The methods of supervised training are those used when labeled samples are available. Among the most popular models are the feed-forward networks, trained under supervision with the back-propagation algorithm. For the present work we have used variations of these basic models, which we describe briefly on the next sections.

5.3 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 [4] 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 [9].

5.4 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 be applied to any network as long as its weights, 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. At each training epoch, if the performance decreases toward the goal, then the learning rate is increased. If the performance increases, the learning rate is adjusted by a decrement factor and the change, which increased the performance, is not made [10].

The training stops when any of these conditions occurs: 1) The maximum number of epochs (loops) is reached, 2) The maximum amount of time has been exceeded, or 3) The performance has been minimized to the goal.

5.4.1 Hybrid System

The hybrid system was designed to train the Input Delay Neural Network with the best features selected from the input vectors. To perform this selection, we apply Evolutionary Strategies, which use real numbers for coding the individuals. The system works as follows.

Suppose we have an original matrix of size s x q, where s is the number of features that each sample has and q is the number of samples available, and we want to select the best features of that matrix to obtain a smaller matrix. For doing so, a population of n individuals is initialized, each having a length of m; these individuals represent n matrices with m rows, and q columns (Fig. 5). Each row corresponds to m random numbers that go from 1 to s.

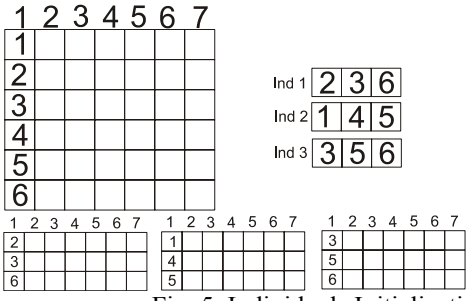

Fig. 5. Individuals Initialization

Once the matrices are obtained, n neural networks are initialized, and we train each one with one matrix, at the end of training each network, we measure the efficiency by using confusion matrices. Once all the results are obtained, we select the matrices yielding the best results (Fig. 6).

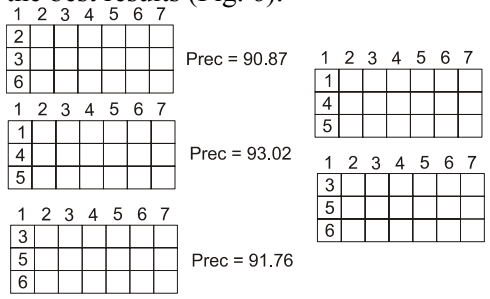

Fig.6. Selection of best individuals.

After the best matrices are selected, they are sorted and ordered from the highest to the lowest value. Next we apply tournament to them, where we generate n random numbers that go from 0 to the number of selected matrices. In Fig. 7 we show that only two matrices where selected, so we generate 3 random numbers from 0 to 2. Since there is no matrix with a 0 index, all 0s randomly generated become 1. As a result the number 1 has twice the probability to be randomly generated than any of the other indexes, which is seen as a reward for best efficiency to the matrix in position 1, getting the highest probabilities to be chosen.

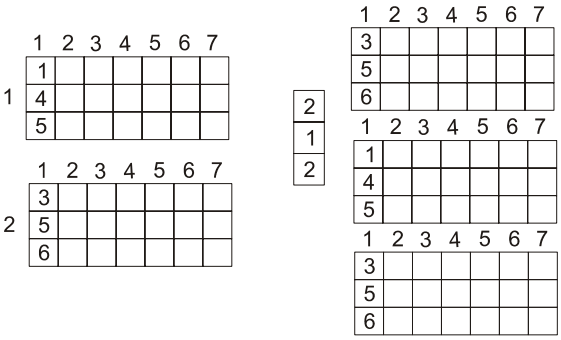

Fig. 7. Generation of a new population with the best individuals.

Once the new generation of individual has been generated, they suffer a random mutation, in each generation. First we choose a mutation factor (we used a mutation rate of 0.2), and when a new descendant is born, we generate a random number that goes from 0 to 1. If it's smaller than 0.2 the individual is mutated, if it's larger, we pass the individual as it is. When the individual is selected to be mutated, we generate a random number that go from 1 to m, this to select which chromosome will be mutated. After we have this position, another random number from 1 to s (e.g. 1 – 6) is generated, which represents a new feature to be selected. When the individual is not selected to be mutated, it goes to the next generation to compete with the others as an exact copy of its parent. This process is repeated for a given number of generations stated by the user.

6 System Implementation for the Cry Classification

On the first stage, the infant's cries are collected by recordings obtained directly from doctors of the Mexican National Institute of the Human Communication (INCH), the Mexican Institute of Social Security (IMSS, Puebla), and by the Voice Processing Group (GPV), from the Universidad de Oriente, in Santiago de Cuba, Cuba.

The cry samples are labeled with the kind of cry that the collector mentions at the end of each cry recording. Later, using the methodology learned from the GPV, we segmented the infant cry into *cry units*. One cry unit consists of a respiration cry from one sample. In other words, we only keep the part of the full sound wave recording where there exists cry information. Each cry unit is segmented in 0.4 seconds; these segments are then labeled with a previously established code, and each one represents one sample. In this way, for the present experiments, we have one corpus made out of 180 samples from *normal* babies, 157 from hypo acoustics (deaf), and 164 with asphyxia. We also have another corpus composed by 154 samples of *normal* babies and 139 samples from pathological cries, both from Cuban babies. The two corpuses were used for separate experiments, as it is later explained. On the next step the samples are processed one by one extracting its LPC and MFCC acoustic features, this process is done with the freeware program Praat 4.2 [11]. The acoustic features are extracted as follows: for each segment we extract 16 coefficients for every 50 milliseconds, generating vectors with 112 features for each sample. The evolutionary algorithm was designed and programmed using Matlab 6.5 R13, the neural network and the training algorithm where implemented using the Matlab's Neural Network Toolbox.

In order to compare the behavior of our proposed hybrid system, we made a set of experiments where the original input vectors were reduced to 50 components by means of Principal Component Analysis (PCA). When we use evolutionary strategies for the acoustic features selection, we search for the best 50 features. In this way, the neural network's architecture consists of a 50 nodes input layer, a 20 nodes hidden layer (40% less nodes than the input layer) and an output layer of 2 or 3 nodes depending on the training corpus in use. The implemented system is totally adaptable, no changes have to be made to the source code to experiment with the Mexican or Cuban corpuses.

To perform the experiments we first separate 25 samples from each class, taking the rest of them to train the system. The 25 samples that we set apart are from babies not used for training, they will be used to test the system. In this way, we try to test the recognition process, as close as possible, under real situations.

From the remaining of the vector sets of each class we first select the acoustic features, and then train the networks. The training is done up to 500 epochs or until a 1×10^{-8} error is reached. Once the network is trained, we test it using the 25 samples from each class separated for this purpose. The recognition precision is shown with the corresponding confusion matrices.

7 Experimental Results

We experimented first with the simple neural network and later with our hybrid system to compare the obtained results. In these experiments we use the same input parameters, in other words, 50 features input vectors and 1 time delay unit.

In the case of the simple neural network system, we perform three experiments and choose the best result. As for the hybrid system, to search for the best solution in a multi solutions space, only one experiment is done. We do so because we use 20 individuals as the initial population, 20 generations to perform the features search, and the size of the individuals was of 50 chromosomes. With all these input parameters, there are 400 different training processes needed, which takes much more time to perform each experiment. The results are shown in Tables 1 and 2, where the first column corresponds to the results obtained from vectors reduced with PCA, and the second one with vectors reduced by selecting features with the evolutionary strategy presented. In both cases, the same kind of input delay neural network was utilized. The only difference being that in the first case the reduction of vectors is done before any processing by the neural network. While in the second case, feature selection is made concurrently to the neural network training. On this basis, we are presenting our model as an evolutionary-neural hybrid system.

Table. 1. Results with the Mexican samples.

	Neural System	Hybrid System
MFCC	93.33%	94.66%
LPC	48%	62.66%

Table. 2. Results with the Cuban samples

	Neural System	Hybrid System
MFCC	88.96%	90.68%
LPC	81.03%	82.75%

8 CONCLUSIONS AND FUTURE WORKS

The application of feature selection methods, on different kinds of pattern recognition tasks, has become a viable alternative tool. Particularly for those tasks which have to deal with input vectors of large dimensions. As we have shown, the use of evolutionary strategies for the selection of acoustic features, in the infant cry classification problem, has allowed us, not only to work with reduced vectors without losing classification accuracy, but also to improve the results compared to those obtained when applying PCA. On the other side, from the results shown, we can conclude that the best acoustic features to classify infant cry, with the presented hybrid system, are the MFCC.

For future works, in order to improve the performance of the described evolutionary-neural system, we are planning to do more experiments using other neural network configurations, different number of individuals, different number of generations and different individual size. We are also looking for adequate models to dynamically optimize the parameters of the hybrid model, in order to adapt it to any type of pattern recognition application.

As for the automatic infant cry recognition problem, we will continue experimenting with some different types of hybrid systems. And once we can be sure that our system is robust enough to identify the mentioned pathologies, we will try to increment our infant cry corpus with some other kinds of pathologies. Particularly with the type of pathologies related to the central nervous system.

Acknowledgments

This work is part of a project that is being financed by CONACYT-Mexico (37914-A).

REFERENCES

[1] Orozco Garcia, J., Reyes Garcia, C.A. (2003), Mel-Frequency Cepstrum coefficients Extraction from Infant Cry for Classification of Normal and Pathological Cry with Feed-forward Neural Networks, ESANN 2003, Bruges, Belgium.

[2] Marco Petroni, Alfred S. Malowany, C. Celeste Johnston, Bonnie J. Stevens, (1995). Identification of pain from infant cry vocalizations using artificial neural networks (ANNs), The International Society for Optical Engineering. Volume 2492. Part two of two. Paper #: 2492-79.

[3] Ekkel, T, (2002). "Neural Network-Based Classification of Cries from Infants Suffering from Hypoxia-Related CNS Damage", Tesis de Maestría. University of Twente, The Netherlands.

[4] 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 Católica del Norte, Chile, 1999, pp 24-29.

[5] Gold, B., Morgan, N. (2000), Speech and Audio Signal Processing. Processing and perception of speech and music. John Wiley & Sons, Inc.

[6] Santo Orcero, David. Estrategias evolutivas, http://www.orcero.org/irbis/disertacion/node217.html. 2004.

[7] Hussain, Talib S., An Introduction to Evolutionary Computation, Department of Computing and Information Science Queens University, Kingston, Ont. K7L 3N6. 1998.

[8] DARPA Neural Network Study, AFCEA International Press, 1988, p. 60.

[9] Limin Fu., Neural Networks in Computer Intelligence. McGraw-Hill International Editions, Computer Science Series, 1994.

[10] Manual Neural Network Toolbox, Matlab V.6.0.8, Developed by MathWoks, Inc.

[11] 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.

[12] Reyes-García C. A. & Reyes-Galaviz O. F., Classification of Infant Crying to Identify Pathologies in Recently Born Babies with ANFIS. ICCHP Transactions, Lecture Notes in Artificial Intelligence, Springer-Verlag. 2004.