A Combined Classifier of Cry Units with New Acoustic Attributes

Sergio Cano¹, Israel Suaste², Daniel Escobedo¹,
Carlos A. Reyes-García², and Taco Ekkel³

¹ CENPIS, University of Oriente, Ave Las Américas 90900 Stgo de Cuba, Cuba

² INAOE, Carretera de Cholula s-n, Puebla, Mexico,

³ Faculty of Informatics, University of Twente, The Netherlands

{scano, danielle}@fie.uo.edu.cu,
{kargaxxi, isuaste}@inaoe.mx, taco@utwente.nl

Abstract. The present work proposes a combined classifier of infant cry units that links in a single structure two focuses: a threshold-based classification and ANN-based classification. The threshold-based classifier considers 4 new acoustic features: *stridor*, *melody*, *voicedness*, *shifts*, that show properly their robustness in front of alterations of the acoustics of infant cry concerned with the presence of some diseases. In order to satisfy the automatic estimation their practical implementations are also considered. The ANN-based classifier consists in a feed-forward network using the method of Scale Gradient Conjugate (MSGC) as learning algorithm and the MFCCs as input vectors to the net. Each focus or classification stage gives in the exit one indicator (FN1 and FN2) that generates to the output a decision on two classes with gradation (*normal*, *moderately-pathologic* and *pathologic*). The results demonstrate the potentiality of these types of combined classifiers when the advantages of each focus in particular are properly emphasized

Keywords: Cry classification, pattern recognition, neural network.

1 Introduction

The problem of cry classification has been recently on the centre of many research efforts because of the avalanche of new focuses in the soft-computing area like *the artificial neural networks (ANN's)*, *genetic algorithms*, *evolutionary computing* and *fuzzy logic* [1-3]. The works of Petroni and Schonweiller using supervised and not supervised ANN's respectively demonstrated the robustness of these models in the detection of abnormalities like deafness in newborns and their ability to classify different cry types. The Group of Speech Processing (GPV) in the University of Oriente developed several experiences with different ANN architectures (MLP, SOM, RBF networks) obtaining good results for the classification of 4 control groups of CNS pathologies (linked with hypoxia: *hypoxia*, *hypoxia with aggravating factors*, *hyperbilirubinaemia* and *delay on intra-uterine growth*). In the 2000 the Mexican group of Cry Analysis of the INAOE began interesting experiences using

neural networks with *fuzzy logic* and *genetic algorithms*. Nevertheless until now no cry-based clinical routines have been achieved for diagnostic purpose, at least in the reported scientific literature. Probably the above-mentioned could be concerned with: the lack of a wide corpus of data to train the learning structures, the inconsistency of the soft-computing algorithms in order to generalize all the patterns learned in the training tests, and the necessity to measure diverse parameters in the acoustics of the cry connected to the infant's neurophysiologic status. Because of which research efforts looking for new approaches in the cry classification continue still being developed.

A current trend is looking for combinations or hybrids of classifiers that could empower the advantages of each one (e.g. genetic-fuzzy approach) [2, 4]. The paper presents a new proposal that combines a traditional focus as the threshold-based classification and the focus using ANN-based classification, resulting in a combined system which displays an abnormality index (with gradation) as output. Moreover the combination of 4 acoustic features of the cry signal is also implemented. Those features have been reported in the bibliography as moderate indicators of certain disorders [5-8] but in this paper, in a combined environment they've shown relevant performances in front of the presence of abnormal cry.

2 Methods and Materials

Despite of the four parameters above mentioned are well known and reported in the literature, they have received insufficient attention according to their potentialities to characterize the neurophysiologic status of the child, raising the attention on other more robust ones as fundamental frequency, formants, latency, etc. Maybe the lack of models or formulas for their automatic calculation has influenced on it.

First, It becomes indispensable to manage some concepts related to these four acoustic parameters:

Fundamental Frequency (F0)¹: it is the lowest frequency component of a complex tone and is the determinant of the pitch that one hears. In infant cries, the fundamental frequency mostly resides in the range of 400 to 600 Hz [8].

Voicedness is defined as being the ratio of the amount of periodic sound versus the amount of noise. The higher the voicedness, the weaker the noise component in comparison to the periodic sound. Hence, the higher the voicedness, the more clearly one could audibly determine the pitch of the sound. Synonyms for a low voicedness are *turbulence*, *disphonation*, *unpitched sound* or simply *noise*.

¹ It is known that in a voiced segment when it is passed through a frequency filter of narrow band in the spectrogram the harmonics appear. The first of these harmonics is the Fundamental Frequency, being the rest of harmonics multiples of the first harmonic. This Fundamental Harmonic has the frequency of the Glottal Pulse (The Glottal Pulse is the impulse of air of the glottal stream produced by movements of opening or closing of the vocal cords. The Fundamental Frequency is proportional in inverse mode to these periods and the unity of measurement is the Hertz, Hz). The Fundamental Harmonic tends to determine the intonation and the melodic curve of any voiced utterance.

Melody is defined as the description of the behavior of the fundamental frequency over time, *within one cry unit*. Normally, cry research identifies six types: rising-falling, falling, flat, falling-rising, rising and glottal plosive.

Shift: it's a sudden large change in pitch (fundamental frequency) caused by the vocal cords crashing into 'overdrive' thereby instantly increasing the fundamental frequency to a much higher value. Those sudden changes are a change of melody or intonation seen on a very relative short time that shows the relation between the fundamental frequency with the stability on the larynx control. [5, 8].

Stridor: another form vocal cord hyperfunction. In this case, a rapid increase in air pressure causes the vocal cords to enter a turbulent state resulting in the sudden loss of pitch. This creates a short noisy (voiceless) segment in the sound, often of about the same energy as the voiced regions surrounding it.

2.1 Acoustical Indicators of Disorder

Various studies have been undertaken to discover possible acoustical features in the infant crying sound that can be correlated to a certain state of the central nervous system. Since the goal is to discriminate between normal and abnormal cases, we need to find these features to perform this classification, in which we could view upon the features as 'decision variables'.

Our four parameters considered in this paper have been connected with certain abnormalities according with the international scientific reports [5,8]: (1) *instant changes in the fundamental frequency of four or more octaves occur significantly more often in abnormal cases*; (2) *the melody type in test cases is, more often than in normal cases, of the rising type*, (3) *the number of turbulent noises (stridors) is increased in abnormal cases*, (4) *in general, the number of shifts and stridors tends to be higher in abnormal cases*. Finally, cries of non-healthy infants are, in general, more often considered to be either hyperfunctional (higher, louder and shriller) or hypofunctional (lower, weaker and thinner) [8].

The research goal is to be able to classify a cry sound as originating from either a healthy (normal) infant or an infant suffering from a CNS disorder (based on hypoxia). In order to perform this classification, the sound needs to be expressed in absolute (as opposed to subjective) terms, and therefore variables that describe the sound are needed. Then, based on the values of these variables, a decision can be made. Considering all of the above, the acoustic parameters to be extracted from the crying sound in order to measure all of the aforementioned indicators are:

- *fundamental frequency over time (just to estimate the melody pattern)*
- *voicedness over time*
- *number of occurrences of shift*
- *number of occurrences of stridor*
- *melody type*

2.2 Procedures to Estimate the Acoustic Features

Voicedness

To extract a measure of 'voicedness' from the sound data, information from the spectral domain can be used. When the data consists primarily of noise, there are few

peaks in the spectral domain of a signal at that time point. However, when the sound consists mostly of periodic sound, clear peaks can be distinguished. To get a measure of this difference, the spectral information is first smoothed over a few frames. This has the effect of a low-pass filter on the frequency data, cancelling out most variability due to white noise. After this, a vector is created containing the differences from one frequency component to the next, effectively creating a numerical derivative. Since peaks in the voiced data create larger differences than the relatively 'flat' data from the noise, the sum of absolute numerical derivatives results in a relatively large value for a voiced region and a small one for a noisy region. The voicedness v is now calculated from an array of amplitudes a (where the indices correspond to frequencies) with length N :

$$v = \sum_{i=1}^{N-1} |a_{i+1} - a_i| \quad (1)$$

Melody

When we visually inspect one cry unit in the spectrogram, coupled with aural inspection the sound, we can determine which melody type is applicable to the unit. For automatic extraction however, things are a bit more complicated. First, there is the fact that the pitch detection is not flawless. If we were to take, for example, only the first frequency value as *start frequency*, a measurement error in this one value could seriously compromise correct detection. To cope with this, a scheme is devised in which all frequency values of one cry unit are pre-processed by multiplying the data within a segment with smoothing windows, as shown in Fig. 1. In doing so, the influence of a single measurement error is greatly reduced while preserving locally relevant information. For smoothing, the frequency values F_i are convolved with a Hanning window to produce a local average, yielding the following equation for calculating the averaged midpoint frequency F_m , where M is the total number of frequency values in the cry unit:

$$F_m = \frac{\sum_{i=1}^M \frac{1}{2} F_i \left[1 - \cos\left(\frac{2\pi}{M} i\right) \right]}{\int_{i=1}^M \frac{1}{2} \left(1 - \cos\left(\frac{2\pi}{M} i\right) \right)} \quad (2)$$

It results in three values: an average frequency representing the beginning of the cry (start), one representing the center of the cry (midpoint), and one representing the end of the cry (end). Secondly, the decision boundaries for the melody type have to be defined in order to answer the question under what numerical situation a certain set of values can be assigned to a certain category. Only the occurrence of a rising melody type will be used in classification. This is considered a valid strategy since, in earlier work, it was concluded that occurrence of the rising melody type was allegedly greater in disorder cases, but no similar significance was found for other melody types [10].

Stridor

Since both voicedness and the energy are calculated over time, stridor can be detected by using the following rule: *when the voicedness suddenly drops within an area of high energy, one occurrence of stridor is marked*. Tangibly, the following thresholds were selected: when the voicedness drops to less than 30% of its maximum

Fig. 1. Windowing procedures used for melody detection

while the energy level remains above -35dB, one occurrence of stridor is marked [5]. An exception is made at the end of a cry unit, because the voicedness of the cry usually drops sharply a little bit before the cry stops. This is a normal consequence of stopping the air stream and it should therefore not be considered a vocal cord hyperfunction. Using this algorithm on the real life data, the number of stridor occurrences per cry unit is usually zero, and sometimes one or two.

Shift

In detecting shift, the definition as used by the GPV in earlier work is followed [9]: when the fundamental frequency instantly (i.e., from one frame to the next) rises four octaves or more, the presence of a shift is marked. One octave represents a doubling of the fundamental frequency, so a stridor is marked when, from one frame to the next,

$$\frac{F_{o_{current}}}{F_{o_{previous}}} > 4$$

2.3 Estimation of Mel Frequency Cepstral Coefficients (MFCC)

The well-known Mel cepstral coefficients were selected to build up the input vectors for the ANN architecture. They have been used with high efficiency in recent works concerned with cry classification. [1,11] The Mel scale filter bank is a series of L triangular band pass filtering believed to occur in the auditory system (corresponding to series of band pass filters with constant bandwidth and spacing on a Mel frequency scale as you see in Fig. 2.

As it is also represented in Fig. 2 the MFCC's can be computed by: (1) generating small segments from the signal, (2) calculating the DFT for each segment, (3) the spectrum is then converted into a logarithmic scale, (4) the scale is transformed into a soft Mel spectrum, (5) the discrete cosine transform DCT is calculated.

Fig. 2. The Triangle-filters use for MFCC computation and the steps required for computing the cepstrum of a cry signal

2.4 The Use of an Artificial Neural Network

The use of ANN's has been a great impact in the development of several research areas like computer vision, autonomous vehicle, pattern recognition, connected-speech synthesis and more recently into the classification of cry units [3-4, 11-12]. The ANN used in this work is shown in Fig. 3. It is a Feed-Forward network in which $x1, x2, ..., xn$ represent the acoustic features of signals and $y1, y2, ..., yn$ the n classes to be identified. This kind of supervised ANN has been also used in cry classification with succeed [11-12].

For an efficient learning of the classification patterns one important aspect to be considered is the training method of the ANN. An optimal learning supposes the minimization of the error function which depends on the weights. For this purposes some algorithms are based on the gradient descendent. Cry data are not static, and any cry sample at any instance in time is dependent on crying patterns before and after

Fig. 3. A Feed-Forward Artificial Neural network

that instance in time. A common flaw in the traditional Back-Propagation algorithm is that it does not take this into account. In the paper we choose the *Method of Scaled Conjugate Gradient* (MSCG) [13].

The MSCG finds out a direction to a new point and then decide what so far to go in that direction (*step size*) using 2nd order information

$$E_{qw}(y) = W(w) + E'(w)^T y + \frac{1}{2} y^T E''(w) y \quad (3)$$

where y is a point in the weight vector. The 2nd order technique uses second derivatives of the objective function and, generally it finds a local minimum but with a high computational cost. The algorithm MSGC shows a linear convergence accentuated in most of the problems [13]. Using a mechanism to decide what so far to go in a specified direction MSGC avoids a time of consumption in the linear learning iteration, which makes the algorithm faster than other second order algorithms.

The soft tools used in this experience were: BDLlanto database with 12 seconds-cry recordings of Cuban children, BPVOZ soft-package, PCVOX and *praat* software for the acoustic signal processing. The ANN implementation (including the MSGC algorithm) was done with *Neural Network Toolbox of Matlab* v. 6.0. [14-15]. The Principal Component Analysis (PCA) was implemented using *Matlab* v 6.0.

3 Results and Discussion

Starting from the primary information in BDLlanto database (32 cases: 16 healthy children and 16 pathological children) a segmentation process was developed to generate the cry units being obtained 73 healthy (normal) cry units and 68 pathological cry units (relative to hypoxia). 58 cry units were chosen (for each class) for training and 10 for classification. The segmentation stage was semi-automatic combining a *begin/end detection* (based on function energy and zero-crossing rate) and a manual correction to reduce the negative effect of considering inappropriate sections within the cry unit. From the cry units obtained from database a parameter estimation for every cry unit is done, following two possible ways:

(a) *estimation of 4-acoustic features for the threshold-based classifier:*

The estimated feature is then compared with the normal threshold values associated to each one of the 4 selected parameters, generating to the exit an index FN1 with the following gradation:

- FN1: 0.25 for 1 parameter altered ²
- 0.5 for 2 parameters altered
- 0.75 for 3 parameters altered
- 1.0 for 4 parameters altered
- 0 for no one parameter altered (normality index)

(b) *estimation of MFCC's for the ANN-based classifier.*

500 MFCC's were computed for each generated cry unit (16 Mel-cepstral coefficients were estimated for each 50 milliseconds frame). The original input vectors were reduced to 50 components by means of Principal Component Analysis

² One parameter altered means that it is out of the threshold boundary for *normality*.

(PCA). Then the input vector to the feed-forward ANN was presented, with the following structure: 50 nodes in the input layer, 15 nodes for the hidden layer and finally 2 nodes for the output layer.

To detect the cry type in the newborn the output values of the net are analyzed. The output values of the net are coded between 0 and 1. If the value of the output node 1 is bigger than the value of the output node 2 the sample is assigned to the class "normal" (N) generating a FN2 index equal to 0, otherwise it is assigned to the class "pathologic" (P) generating a FN2 index equal to 1.

Finally both FN1 and FN2 indexes are processed in a decision block ($D = \frac{FN1 + FN2}{2}$) resulting in two classes-based decisions with 3 qualitative levels:

Normal $D \leq 0.5$

Moderately- pathologic $D = 0.75$

Pathologic $D = 1.0$

Tables 1, 2, 3 and 4 display the results from the threshold-based classifier, ANN-based classifier and the Combined Classifier respectively. From these results some comments can be done:

- The present performance is higher than other similar works [1-2]
- In Table 4 the gradation of the D index let physicians to use properly the output of the cry classifier in order to compare and to evaluate its "possible meaning" in front of the results from the neurophysiologic evaluation of the newborn (how much abnormal the infant cry is from the acoustical point of view and its "weight" for diagnostic purpose).
- The need to include more acoustic features in cry classifier for better classification rates proposed and argued by Schonweiller in 1996 [2], is well demonstrated here. As you see the FN1 index displayed a strong correlation between the altered status of acoustic feature and the pathologic status of cry unit.
- Of particular interest was the fact that the only two cry units misclassified as normal obtained a FN1 equal to 0.75 (abnormal for the threshold-based classifier), so both outputs from the classifiers also offer valuable information to be considered by the specialists (see table 2).
- The higher frequency of altered parameters for pathological cry units seems to be representative of the robustness of the four acoustic features used as detectors of abnormality. (see table 3)

Table 1. Results from the Threshold-based classifier

		FN1 Index				
		0	0.25	0.5	0.75	1.0
<i>Normal</i>	10	9	1	-	-	-
<i>Pathologic</i>	10	-	2	5	2	1

Table 2. Frequency of altered parameters for both classes (p/n)

	Altered Parameter			
	Stridor	Voicedness	Melody	Shift
<i>Normal</i> (10)	-	-	-	1
<i>Pathologic</i> (10)	6	5	5	6

Table 3. Results from the ANN-based classifier

	Cry Units	Confusion Matrix		% Classification
		Normal	Pathologic	
<i>Normal</i>	10	7	3	70%
<i>Patolog.</i>	10	2	8	80%
<i>Total</i>	20			75%

Table 4. Results from the Combined Classifier

		Confusion Matrix		D index			% Classification
		N	P	$X \leq 0.5$	$0.5 < x \leq 0.75$	$0.75 < x$	
<i>Normal</i>	10	10	0	10	0	0	100
<i>Pathologic</i>	10	2	8	2	7	1	80
<i>Total</i>	20			12	7	1	90

4 Conclusions

In the presented work we have developed a new combined-cry classifier which uses two focuses for cry classification (a threshold-based classifier and an ANN-based classifier) with satisfactory results. Moreover four acoustic features as *shift*, *voicedness*, *melody* and *stridors* were estimated and efficiently combined, showing their potentialities to detect abnormal behaviour related to the neurophysiologic status of the newborn. Both output indexes FN1 and FN2 offer also valuable information for specialists when they analyze them together or in separate environment.

Another interesting aspect was concerned with the successful use of *cry unit* as the basic processing unit for cry classification validating the assessment done by Ekkel in 2000 [16]. The asseveration made by Schonweiller in [2] about to include other parameters in the domain frequency it is also supported here.

Further tests should focus on using different acoustic features and different soft-computing paradigms. Also the use of hybrid structures for cry classifiers seems to be a promise alternative.

Acknowledgments

This work is in part supported by National Council for Science and Technology CONACYT-Mexico under project (C01-46753).

References

1. Petroni, M., Malowany, A., Johnston, C., Stevens, B. "*Identification of Pain from Infant Cry Vocalizations Using Artificial Neural Networks (ANN's)*". The International Infant Cry Research Group. Applications and Science of Artificial Neural Networks. The International Society for Optical Engineering. Volumen 2492. 1995. pp.729-738.
2. Schonweiler, R. et al. "Neuronal networks and self-organizing maps: new computer techniques in the acoustic evaluation of the infant cry". Int. Journal of Pediatric Otorhinolaryngology. Elsevier Science, 38, 1996, pp. 1-11
3. Suaste, I, Reyes-Galaviz, O., Dias, A., Reyes, C. A Fuzzy Relational Neural Network for Pattern Classification. A. Sanfeliu et al (Eds.): CIARP 2004, LNCS 3287, pp. 358-365, 2004.
4. Alonso, L. et al. *Reconocimiento de Patrones con Redes Neuronales*. Ed. L. Alonso, Editorial Imprenta Catedral, 2001, pp 37-56
5. Wasz-Höckert, O., Lind, J., Vuorenkoski, V., Partanen, T., Valanne, E. "*The Infant Cry: A Spectrographic and Auditory Analysis*". Spastics International Medical Publication, Lavenham, U.K., 1968.
6. Hurlock, E.B. "*Child Development*". Nueva York. McGraw-Hill. 1950.
7. Zeskind P.S., Lester B. "Acoustic features and auditory perceptions of the Cries of Newborns with Prenatal Complications". Child development, pp. 580-589. 1978.
8. Michelsson, K. "Sound Spectrographic cry analysis of normal and abnormal newborns". Folia Phoniatra, 28, 1982, pp. 161-173
9. Escobedo, D., Cano, S., Coello, E., Regueiferos, L., Capdevila, L.: Rising Shift of Pitch Frequency in the Infant Cry of Some Pathologic Cases. 2nd Int. Conf. MAVEBA 2001, Firenze, Italy, 2001-
10. Cano, S. et al. "The Spectral Analysis of Infant Cry: An Initial Approximation". Proceedings of EUROSPEECH'95 (sponsored by ESCA & IEEE), Madrid, Sept 18-21, 1995
11. Orozco, J., Reyes, C. "Extracción y Análisis de Características Acústicas del Llanto de Bebés para su Reconocimiento Automático Basado en Redes Neuronales". Tesis de Maestría, INAOE, Puebla, Mex., 2002.
12. Cano, D., et al. "Análisis preliminar de los resultados de una clasificación de unidades de llanto según tres arquitecturas de redes neuronales". Memorias de TELECOM'2002 (en CD-Rom). Centro de Convenciones HEREDIA, Julio 2002.
13. Moller. "A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning". Neural Networks 6 (4). 1993. pp. 525-533.
14. Boersma, P., Weenink, D., Praat v. 4.0.27. "Sistema para hacer fonética por computadora". University of Amsterdam. The Netherlands. 1992-2002.
15. The MathWorks, Inc., MATLAB®. "Neural Network Toolbox". Version 6.0.0.88. 1984 - 2000.
16. Ekkel, T. "Neural Network-Based Classification of Cries from Infants Suffering from Hypoxia-Related CNS Damage". Tesis de Maestría. University of Twente. The Netherlands, 2002.