Exploring feature extraction methods for infant mood classification

Leandro D. Vignolo *, Enrique M. Albornoz and César E. Martínez
Research institute for signals, systems and computational intelligence, sinc(i),
Universidad Nacional del Litoral - CONICET, Santa Fe, Argentina
E-mails: ldvignolo@sinc.unl.edu.ar, emalbornoz@sinc.unl.edu.ar, cmartinez@sinc.unl.edu.ar

Abstract. Speaker state recognition is an important issue to understand the human behaviour and to achieve more comprehensive speech interactive systems, and therefore has received much attention in recent years. This work addresses the automatic classification of three types of child emotions in vocalisations: neutral mood, fussing (negative mood) and crying (negative mood). Speech, in a broad sense, contains a lot of para-linguistic information that can be revealed by means of different methods for feature extraction and, in this case, these would be useful for mood detection. Here, several set of features are proposed, combined and compared with state-of-art characteristics used for speech-related tasks, and these are based on spectral information, bio-inspired ear model, auditory sparse representations with dictionaries, optimised wavelet coefficients and optimised filter bank for cepstral representation. All the experiments were performed using the Extreme Learning Machines as classifier because it is a state-of-art classifier and to achieve comparable results. The results show that by means of the proposed feature extraction methods it is possible to improve the performance provided by the baseline features. Also, different combinations of the developed feature sets were studied in order to further exploit their properties.

Keywords: mood classification, crying detection, sparse representations, filter bank optimisation, spectral features, bio-inspired ear model, wavelet packets

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

1. Introduction

People are able to interpret the implicit and explicit information present in human communications in order to arrive at diverse judgements about the messages and the speaker states [11, 25]. The scientific community use the concept of *speaker state* in different scopes, consequently, the word "state" can refer to psychological states, emotional states, sleepiness degrees, specific illness states, among others. Recognition of diverse speaker states has become a multi-disciplinary research area that has drawn great interest over the last years [15, 45, 46]. Specifically, crying is an important communication tool for infants to express their emotional states and psychological needs [19]. Since infant may cry for a variety of reasons, parents and childcare specialists need to be able to distinguish between different types of cries through their auditive perceptions. However, this requires experience and this can be subjective from one person to another. Also, it has been demonstrated that the experienced subjects are often not able to explain the basis of such skills [19]. This motivates the work on the development of automatic tools for the analysis and recognition of infant cry applicable to real life. For example, the use of spectrogram representations from short audio segments as input to convolutional neural network was proposed in [50]. Also, the combination of a dynamic neural network and a multilayer perceptron was proposed, using as input a set low-level features and a large set of statistical functionals applied to the same features, respectively [29].

Many approaches have been proposed to deal with the problem of feature extraction from audio signals, and many of them are focused on aspects like human auditory perception. Among them, the MFCC (Mel frequency cepstral coefficients) are one of the most widespread features for speech processing [18]. Some works use them for speech recognition [51], speaker identification [4], emotional state recognition [33], or spoken language classification [8]. The MFCC features have also been used for recognition of pathologies in recently born babies through their crying [40], for the analysis of infant cry with hypothyroidism [62] and

*Corresponding author. E-mail: ldvignolo@sinc.unl.edu.ar.

for classification of normal and pathological cry [24]. Also, the use of MFCC features was proposed for cry signal segmentation and boundary detection of expiratory and inspiratory episodes [1]. The MFCC features are based on the mel filter bank, which mimics the frequency response in the human ear by means of a perceptual scale of pitches judged by listeners [18]. However, as the physiology of human perception is not yet fully understood, it is not possible to say that this is the optimal filter bank. Moreover, the "optimal" will be necessarily depending on the application and then, it is not clear that one unique filter bank would be able to enhance the information more relevant for any task. This has motivated the development of many approaches for tuning the filter bank in order to obtain better representations [2, 30, 32]. For example, a weighting function based on the harmonic structure was proposed for improving the robustness of MFCC [26]. In the same way, other ideas for tuning the parameters of the mel filter bank have been introduced [59, 61]. In [48], a scheme for determining filter bandwidth was presented, showing speech recognition improvements with respect to traditional features. Also, auditory features based on Gammatone filters were developed for robust speech recognition [47].

A common approach that has been used for many different machine learning problems is to introduce learning in the pre-processing step for producing optimised features [38, 52]. That is the case in [49], where a deep learning approach was used to optimise the features used in an end-to-end approach. The versatility of genetic algorithms has motivated many approaches for feature selection [39, 54], like the optimisation of wavelet decompositions for speech recognition [53]. Also, many other strategies for developing optimised representation for speech related tasks have been presented [55, 56]. Evolutionary approaches have also shown success for the development of new features for stressed speech classification [7]. Although, to our knowledge, the evolutionary optimisation of representations for the cry recognition task has not been explored.

The use of biologically inspired, feature extraction methods has improved the performance of artificial systems that try to emulate some aspects of human communication. Based on the biological time-frequency analysis the inner ear carries out, representations of speech beyond the cochlea have been proposed which allow the estimation of *auditory spectrograms* [17]. Then, the discharge patterns of auditory nerves can be modelled using the notion of *spectro-temporal receptive fields* (STRF), defined as the optimal linear filter that convert a time-varying stimulus into the firing rate of an auditory cortical neuron, so that it responds with the largest possible activation [22]. Using two-dimensional discrete dictionaries, an approximated cortical representation can be established by means of sparse representations [16, 20, 42], with the meaning of the set of activations that contribute to form a particular pattern from an estimation of a STRF.

In this work, the automatic classification of crying vocalisations (to allow automatic mood monitoring of babies for clinical or home applications [45]) is tackled using different approaches for features extraction based on: spectral information, a bio-inspired ear model, auditory sparse representations with dictionaries and two feature optimisation approaches bases on evolutionary algorithms (EA). The first EA approach is based on the conventional signal processing technique for the classical MFCC and is focused on the optimisation of the filter bank involved in the feature extraction process. The second evolutionary approach, unlike the first one, exploits a non-conventional feature extraction procedure based on wavelet packets [53]. In this approach, an evolutionary algorithm is used for the selection of the most relevant features within a large number of wavelet coefficients. For the spectral features, the Fourier spectrogram and an auditory spectrogram are used. Finally, based on this auditory time-frequency representation, a dictionary of two-dimensional atoms is estimated and the sparse vectors of activations (*approximated cortical representation*) are computed as features.

The organisation of the paper is as follows. Section 2 explains all the proposed methods for speech signal representation used in this work. Also, the database and the classifier are introduced. In Section 3, the results obtained considering different combinations of the proposed feature sets are presented and discussed. Finally, Section 4 summarises the contributions of this paper and outlines future research lines.

2. Feature extraction methods and classifier

This section firstly presents the speech database and the baseline features, and then, our approaches are introduced.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

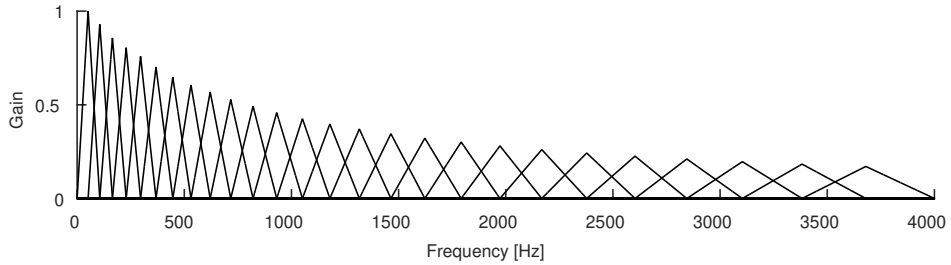

Fig. 1. Filter bank designed based on the mel scale.

2.1. Speech corpus and baseline systems

For the experiments the Cry Recognition In Early Development (CRIED) corpus was used, which is composed of 5587 utterances [45]. The vocalisations were produced by 20 healthy infants (10 male and 10 female), each of which was recorded 7 times. The corpus consists of audio-video recordings, though only audio is considered in this work. The original audio is sampled at 44.1 kHz and was down-sampled to 8 kHz in this work for the filter bank optimisation. This database was made available for the Crying Sub-Challenge of the Interspeech 2018 Computational Paralinguistics ChallengE (ComParE) [45].

The database is split into training and test partitions. The utterances were classified into the following three categories:

- neutral/positive mood vocalisations,
- fussing vocalisations, and
- crying vocalisations.

The categorisation process was done on the basis of audio-video clips by two experts in the field of early speech-language development [36]. In the experiments only audio recordings were considered and, since the labels for the instances of the test partition are not available, cross validation (CV) was performed using the training data.

In order to compare the proposed features with a well known representation, a set of features based on the mel frequency cepstral coefficients [18] was considered as a baseline.

The cepstral analysis is a special case of the homomorphic processing methods and it is applied in speech signal analysis to extract the vocal tract information. Based on the Fourier Transform (FT), the cepstrum is defined as $cc(n) = FT^{-1}\{\log|FT\,\{x(n)\}\,|\}$.

An analysis that combines the cepstrum properties and experimental results about the human perception of pure tones brings out the MFCC representation. A mel is a unit of measure for perceived pitch or frequency of a tone and the mel scale was determined as a mapping between real frequency scale (Hz) and the perceived frequency scale (mel)

$$F_{mel} = 1000 \log_2 \left[1 + \frac{F_{Hz}}{1000}\right]. \quad (1)$$

Based on this scale, n filters with centres equispaced in mels are mapped into linear frequency to obtain a the filter bank shown in Figure 1. Summarising, to obtain the MFCC coefficients, the FT is calculated and this spectrum is filtered by a filter bank in the mel domain [18]. Then, the logs of the powers at each of the mel frequencies are taken. Finally, the FT is replaced by the Cosine Transform in order to simplify the computation and it is used to obtain the MFCC of the list of mel log powers.

The first 17 MFCCs were computed on a time frame basis, using a 20-ms window with 10-ms step. Then, the feature set was obtained by applying a number of functionals (listed on Table 1) on the MFCCs, resulting in 531 attributes. These features are considered because they are widely used in many speaker state recognition tasks.

2.2. Mean of log-spectrum

The Mean of Log-Spectrum (MLS) coefficients is a set of features calculated from spectral data for different frequency bands. They were defined to extract relevant information from speech signals and were firstly used in the analysis and classification of spoken emotions (in clean and noisy conditions [5, 6]). The MLS coefficients are defined as the average of the signal spectrogram

$$S(k) = \frac{1}{N}\sum_{n=1}^{N} \log|v(n,k)|, \quad (2)$$

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

Table 1
Functionals applied to MFCC [21, 46].

quartiles 1-3	mean value of peaks - arithmetic mean
3 inter-quartile ranges	linear regression slope and quadratic error
1 % percentile (≈ min)	quadratic regression a and b and quadratic error
99 % percentile (≈ max)	arithmetic mean, standard deviation
percentile range 1 %-99 %	standard deviation of peak distances
simple moving average	contour is below 25 % range
skewness, kurtosis	contour is above 90 % range
mean of peak distances	contour is rising/falling
mean value of peaks	linear prediction of MFCC contour (coefficients 1-5)
contour centroid	gain of linear prediction

where k is a frequency band, N is the number of frames in the signal and $v(n,k)$ is the discrete Fourier transform of the signal in the frame n. The spectrograms were obtained with Hamming windows of 25 ms, and for 16kHz sampled signals, 200 MLS coefficients corresponding to equally spaced frequency bands are obtained.

2.3. Auditory cortical representation

In neuroscience, it has been established the principle that the brain of an animal adapt its properties (internal configuration) to best describe the statistics of stimuli perceived through its senses [9]. The representation of the sound signal at the cochlear level and auditory cortical areas has been studied as an alternative to classical analysis methods, given its intrinsic selective tuning to relevant natural sound [58]. In [60], a model based on neuro-physiological investigations at various stages of the auditory system was proposed. This model has two consecutive stages: an early auditory spectrogram with the activity of auditory nerve fibres (Figure 2.3), and a model of the primary auditory cortex used to process the spectrogram and find the spectro-temporal receptive fields. The first stage uses a bank of 128 cochlear (bandpass) filters in the range $[0 - 4000]$ Hz, with the central frequency of the filter at location x on the logarithmic frequency axis (in octaves) is defined as $f_x = f_0 2^x(\text{Hz})$, where f_0 is a reference frequency of 1 kHz. This frequency distribution proved to be satisfactory for the discrimination of acoustic clues in speech and further reconstruction of the signals [13].

Here, the bio-inspired representation is used in two different manners which will be presented at next.

Mean of the log-auditory spectrum

In the same way as previously for MLS, we propose to analyze the recordings by means of a related set of features based on the auditory spectrogram. Using the first stage output, a set of features is built using the mean of the log auditory spectrogram (MLSa) [6], as

$$S_a(k) = \frac{1}{N}\sum_{n=1}^{N} \log |a(n,k)|, \tag{3}$$

where k is a frequency band, N is the number of frames in the utterance and $a(n,k)$ is the k-th coefficient obtained by applying the auditory filter bank to the signal in the frame n.

The MLSa was computed using auditory spectrograms calculated for windows of 25 ms without overlapping. In order to obtain the representation of sound in the auditory model, a Matlab implementation of the Neural System Lab auditory model was used[1].

All MLS and MLSa features were computed on a frame by frame basis in order to compute statistics (mean and standard deviation) for each utterance.

In order to reduce the number of features obtained with MLS and MLSa, maintaining the most relevant for this classification problem, a ranking feature selection procedure was performed based on the F-Score measure [12]. The F-Score rates the features based on their discriminative capacity. Given a feature vector FV_k, this score was computed considering the True instances (N_T) and the False instances (N_F) as follows:

[1] Neural Systems Lab., Institutes for Systems Research, UMCP. http://www.isr.umd.edu/Labs/NSL/

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

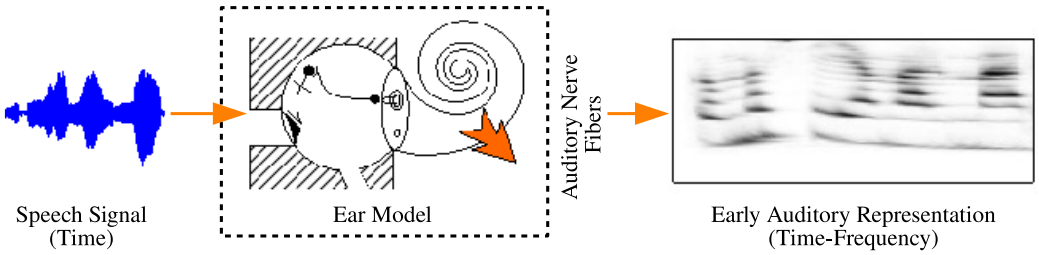

Fig. 2. Scheme of the used auditory model.

$$F(i) = \frac{\left(\bar{x}_i^{(T)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(F)} - \bar{x}_i\right)^2}{\frac{1}{N_T-1}\sum_{j=1}^{N_T}\left(x_{j,i}^{(T)} - \bar{x}_i^{(T)}\right)^2 + \frac{1}{N_F-1}\sum_{j=1}^{N_F}\left(x_{j,i}^{(F)} - \bar{x}_i^{(F)}\right)^2} \tag{4}$$

where $\bar{x}_i$ is the average of the ith feature, $\bar{x}_i{}^{(F)}$ and $\bar{x}_i{}^{(T)}$ are the average False and True instances respectively, and $x_{j,i}$ is the ith feature in the jth instance.

This work proposes the use of MLS and MLSa features separately and also both sets combined. In order to combine the feature sets two approaches were considered. In the first approach the features in each set are ranked separately according to F-Score, and the higher ranked features are kept for each set [MLS+MLSa (Added)]. In the second approach all the MLS and MLSa features are ranked together by F-Score, in order to select the higher ranked features [MLS+MLSa (Combined)].

Sparse representation based on discrete dictionaries

There are different ways of representing a signal using general discrete and finite dictionaries. For the case where the dictionary forms a basis, in particular for the orthonormal or unitary cases, the techniques are quite simple. This is because, among other aspects, the representation is unique. However, in the general case, a signal can have many different representations for the same dictionary. In these cases, it is possible to find a suitable representation if additional criteria are imposed. For our problem, these criteria can be motivated by obtaining a representation with characteristics such as sparseness and independence. Furthermore, it is possible to find an optimal dictionary that resembles biological properties of sensorial systems.

A sparse code represents the information in terms of a small number of descriptors taken from a large set. In numerical terms, this means that the majority of the elements are zero, or 'almost' zero, most of the time [31]. Considering $\mathbf{x} \in \mathbb{R}^{m\times n}$ as sliding patches from the auditory spectrograms, we can have a linear combination of atoms representing the features in the form

$$\mathbf{x} = \Phi\mathbf{a} + \varepsilon, \tag{5}$$

where $\Phi \in \mathbb{R}^{m\times n\times M}$ is the dictionary of M bidimensional atoms, $\mathbf{a} \in \mathbb{R}^M$ is the target representation and ε is the term for additive noise. In the context of this work, $\mathbf{x}$ corresponds to the reconstruction of the time-frequency representation of the speech at the auditory cortex. The atoms in Φ will further be the representation of the important features found at the cortex for each input stimuli. Finally, an estimation of the coefficients $\mathbf{a}$ will be the output of the feature extraction stage proposed. The system (5) can also be seen as a generative model. Following the terminology used in the ICA field, this means that signal $\mathbf{x}$ is generated from a set of sources a_i (in the form of a state vector $\mathbf{a}$) using a mixing matrix Φ and including an additive noise term ε (Gaussian, in most cases).

Although (5) appears very simple, the main problem is that for the most general case all the variables are unknown, thus there can be an infinite number of possible solutions. Even in the noiseless case (when $\varepsilon = 0$) and given Φ, if there are more atoms than the dimension of $\mathbf{x}$ then multiple representations of the signal are possible (over-complete dictionaries). When Φ and $\mathbf{x}$ are known, an interesting way to choose the set of coefficients $\mathbf{a}$ from among all the possible representations, consists of finding those a_i which make the representation as sparse and independent as possible. In order to obtain a sparse representation, a distribution with positive kurtosis can be assumed for each coefficient a_i. Further, assuming the statistical independence of the a_i, the imposed joint *a priori* distribution satisfies

$$P(\vec{a}) = \prod_i P(a_i). \tag{6}$$

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

The basis functions are vectorized as $\Phi = [\vec{\Phi}_1 \dots \vec{\Phi}_M]$ with $\vec{\Phi}_i \in \mathbb{R}^{[mn]\times 1}$. The sparse representation is obtained when the solution is restricted to

$$\min_a \|\mathbf{a}\|_0, \tag{7}$$

where $\|\cdot\|_0$ is the l^0 norm. This is a known NP-complete problem so different approximations were proposed [37].

In order to obtain the representation, two problems have to be jointly solved: the estimation of the sparse representation and the inference of the dictionary of specialized atoms. As the auditory spectrograms are, in fact, images, a positive-valuated representation is desired along with the sum of (only) positive atoms. Thus, the Non-Negative Singular Value Decomposition method (NN-K-SVD) is used [3]. This method also provides the possibility to fix the number of the sparse components to use in the approximation and solves the problem

$$\min_a \|\mathbf{x} - \Phi^L \mathbf{a}\|_2^2 \;\; s.t.\; \mathbf{a} \geqslant 0, \tag{8}$$

where a sub-matrix Φ^L that includes only a selection of the L largest coefficients is used. In the dictionary updating, this matrix is forced to be positive by calculating

$$\min_{\vec{\phi}_k, a^k} \|\mathbf{E}^k - \vec{\phi}_k a^k\|_2^2 \;\; s.t.\; \vec{\phi}_k, \vec{x}^k \geqslant 0, \tag{9}$$

for each one of the k selected coefficients. The error matrix $\mathbf{E}^k$ is the residual between the signal and its approximation, while k-th atom $\vec{\phi}_k$ and its respective activation a^k are updated. Finally, the dictionary itself, and the activation coefficients are calculated from the SVD of $\mathbf{E}^k = \mathbf{U}\Sigma\mathbf{V}^T$, where the atoms and activations are obtained as the rank-one approximation with the first left and right singular vector as $\phi_k = \mathbf{u}_1$ and $a^k = \mathbf{v}_1$.

Once the dictionary is obtained, for comparison purposes two different setups were defined in order to get the representation. The first setup consists on obtaining the mean activation vector for each signal, that is, 1 feature vector from file. The second setup consists on calculate the activations for each sliding window and feed them to the classifiers. Here, the number of patterns extracted from each audio signal depends on the length of the file. This is the so-called *frame level classification*. Both setups use the notion of *non-zero activations per pattern* (NNAPP) as the number of k selected coefficients in the NN-K-SVD algorithm.

2.4. Evolutionary filter bank optimisation

In order to analyse the appropriateness of the mel mapping (Eq. 1) for infant cry recognition, the mean log-spectrum was computed along the frames (30 ms long) for all the training utterances in each class of the CRIED corpus. As it can be observed on top of Figure 3, the plots corresponding to different classes show different peaks at different frequency bands, suggesting that the relevant information is not mainly at low frequency bands.

Also, the first-order difference of the mean log-spectrums were computed, which are shown at the bottom of Figure 3. These plots present peaks at high frequency bands showing different relative energy and shape, which could be useful for classification. Since the mel filter bank (shown on top of Figure 9) prioritizes low frequencies with higher resolution and amplitude, all these remarks suggest that it is not entirely appropriate for this task. This motivates the work in a methodology useful for finding an optimal filter bank for the task at hand.

The proposed optimisation approach, referred to as *Evolutionary Spline Cepstral Coefficients* (ESCCs), is based on an EA to search for the optimal filter bank parameters. In this approach, instead of encoding the filter bank parameters directly, the candidate solutions in the EA use spline functions to shape the filter banks. In this way, the chromosomes (candidate solutions) in the population of the EA hold spline parameters instead of filter bank parameters, which reduces the chromosome size and the search space. With this encoding, the chromosomes within the EA population contain spline parameters instead of filter bank parameters, reducing the size and complexity of the search space. The spline mapping was defined as $y = c(x)$, with $y \in [0, 1]$, and x taking n_f equally spaced values in $(0, 1)$. Then, for a filter bank with n_f filters, value x_i was assigned to filter i, with $i = 1, ..., n_f$. For a given chromosome, the y_i values were computed for each x_i by means of cubic spline interpolation. The chromosomes encoded two splines: one to determine the frequency values corresponding to the position of each triangular filter and another to set the amplitude of each filter.

2.4.1. Optimisation of filter frequency locations

A monotonically increasing spline is used here, which is constrained to $c(0) = 0$ and $c(1) = 1$. Four parameters are set to define the spline I: y_1^I and y_2^I corresponding to fixed values x_1^I and x_2^I, and the derivatives, σ and ρ, at the fixed points $(x = 0, y = 0)$ and

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

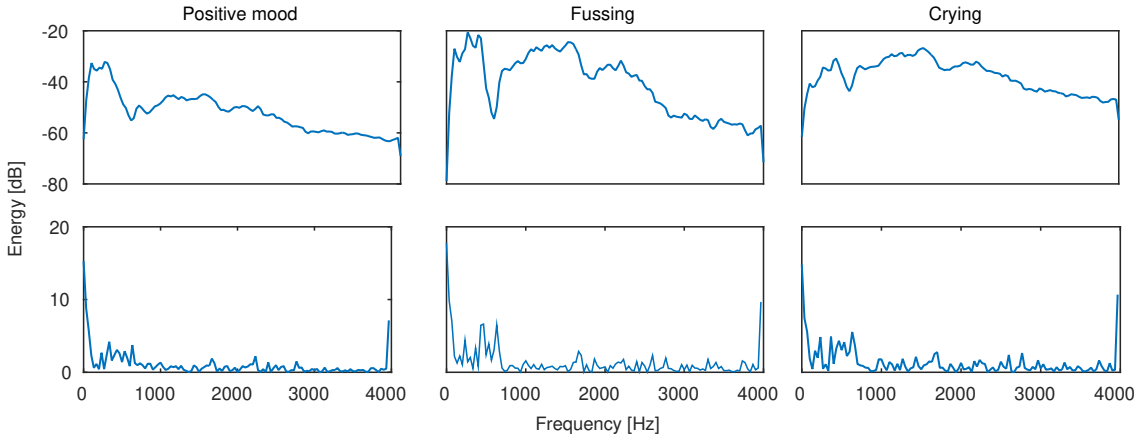

Fig. 3. Mean log-spectrums (top) and first-order difference of mean log-spectrums (bottom) for each of the three classes in the Cry Recognition In Early Development (CRIED) database.

$(x = 1, y = 1)$. Then, parameter y_2^I was obtained as $y_2^I = y_1^I + \delta_{y_2}$, and the parameters actually coded in the chromosomes were y_1^I, δ_{y_2}, σ and ρ. Given a particular chromosome, which set the values for these parameters, the $y[i]$ corresponding to the $x[i]\ \forall\ i = 1, ..., n_f$ were obtained by spline interpolation.

The $y[i]$ values obtained through the spline were then mapped to the frequency range from 0 Hz to $f_s/2$, so the frequency values for the maximum of each of the n_f filters, f_i^c, were obtained as

$$f_i^c = \frac{(y[i] - y_m) f_s}{y_M - y_m}, \tag{10}$$

where y_m and y_M are the spline minimum and maximum values, respectively. Then, the filter spacing was controlled by the slopes of the corresponding points in the spline.

Also a parameter $0 < a < 1$ was defined to limit the range of y_1^I and y_2^I to $[a, 1 - a]$, with the purpose of keeping the splines within $[0, 1]$.

2.4.2. *Optimisation of filter amplitudes*

The spline used for optimising filter amplitudes were restricted to the range $[0, 1]$, but y was free at $x = 0$ and $x = 1$. Therefore, the parameters to be optimised here were the y values y_1^{II}, y_2^{II}, y_3^{II} and y_4^{II}, corresponding to the fixed x values x_1^{II}, x_2^{II}, x_3^{II} and x_4^{II}. These four y_j^{II} were limited to $[0, 1]$. In this manner, n_f interpolation values were obtained to set the amplitude of each filter. This is shown in Figure 4, where the gain of each filter was set according to the value given by spline II at the corresponding points.

2.4.3. *ESCC optimisation process*

Every chromosome in the EA the contains a set of spline parameters that encode a particular filter bank. The search performed by the EA is guided by the classification performance, which is evaluated for each candidate solution. In order to evaluate a candidate solution, the ESCC feature extraction process was performed on the corpus based on the corresponding filter bank (Figure 4). Then, the classifier is trained and tested using the features obtained through this process in order to assign the fitness to the corresponding individual.

The spline codification scheme allowed to reduce the chromosome length from $2n_f$ to the number of spline parameters. Since 26 filters were used, the number of free parameters in the optimisation was reduced from 46 to 8 (4 parameters for each spline). The spline parameters were randomly initialised in the chromosomes using uniform distribution.

After computing the cepstrum from the sequence of output filter bank energies, 15 cepstral coefficients are retained for each frame of analysis. For each of these coefficients, the mean, standard deviation and kurtosis across all frames are computed. Then, a vector of 45 features is obtained for the ESCCs to parameterise each sound file.

Based on previous works, the population size was set to 30 individuals, while crossover and mutation probabilities were set to 0.9 and 0.12, respectively [55, 56]. In this EA, tournament selection and standard one-point crossover methods were used, while the mutation operator was designed to modify splines parameters. The parameters were randomly chosen by the op-

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

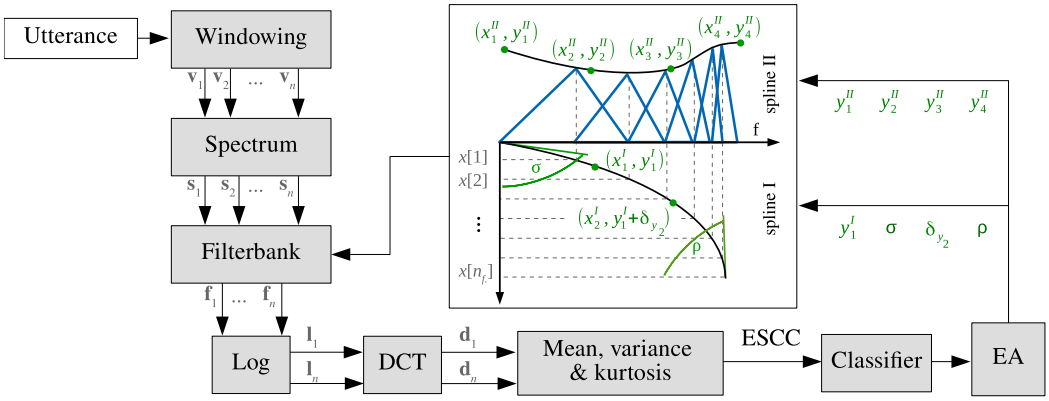

Fig. 4. Schematisation of the optimisation strategy. The output vectors of each block, $\mathbf{s}_i$, $\mathbf{f}_i$, $\mathbf{l}_i$ and $\mathbf{d}_i$, indicate that each window $\mathbf{v}_i$ is processed isolated and, finally, the mean and variance for each coefficient is computed from the $\mathbf{d}_i$ vectors in order to feed the classifier.

erator and the modifications were performed using a uniform random distribution.

2.5. Optimisation of wavelet based features

2.5.1. Wavelets and wavelet packets

Wavelet bases have the property of being localised in both time and frequency, which makes them useful for the analysis of signals with transient and stationary behaviours like speech. Wavelets are functions with zero mean and unitary norm [35], which are translated and scaled in order to obtain the time-frequency atoms. The discretisation of scaling and translation parameters, particularly with scaling factor 2^j, gives the discrete dyadic wavelet transform. The wavelet packets transform (WPT) is implemented by convolving the signal with a pair of quadrature mirror filters (low-pass and high-pass) to decompose the signal into detail and approximation coefficients [35]. Both approximation and detail signals are further decomposed within an iterative process, in which the frequency resolution is increased. Each filtering step is computed as:

$$c_{j+1}^{2r}[m] = \sqrt{2} \sum_{n=-\infty}^{\infty} g[n-2m] c_j^r[n], \tag{11}$$

$$c_{j+1}^{2r+1}[m] = \sqrt{2} \sum_{n=-\infty}^{\infty} h[n-2m] c_j^r[n], \tag{12}$$

where $g[n]$ and $h[n]$ are the impulse responses of the high-pass and low-pass filters associated to the wavelet and scaling functions, respectively, j is the depth of the node and r is an index for the nodes which lay on the same depth. Then, c_j^{2r} is referred to as the approximation of c_{j-1}^r, and c_j^{2r+1} is referred to as the detail. The result offers flexibility for frequency band selection, as shown in Figure 5, providing an over-complete dictionary. The decomposition offered by the WPT allows to analyse a signal in a flexible time-scale plane, in which different sub-trees or nodes can be selected to extract the desired information from the full decomposition. Choosing one among all the possible combinations for a particular application is a challenging problem, which is usually solved by restricting the search to orthogonal basis using diverse criteria [23, 57]. For signal compression, for example, the most common paradigms are based on the *best orthogonal basis* [14] and the *local discriminant basis* [44] algorithms. However, for the classification problem, the convenience of an orthogonal basis has not been proved. Moreover, recent studies conclude that a thorough search within the full decomposition to find redundant representations could provide robustness to obtain better classification performance in noise conditions [53].

2.5.2. Evolutionary wavelet features

In the feature extraction process we used 32 ms windows and the WPT process of filtering and decimation was performed to obtain a decomposition tree of 8 levels, 2048 coefficients and 510 nodes (original signals were down-sampled to 8kHz). In [53] an ad-hoc integration scheme for reducing the search space. This means that each wavelet tree node was divided into groups, according to the integration scheme, in order to compute the energy in each group. Note that only 6 decomposition levels were considered in that integration scheme. In this work a different approach was used, in which the coefficients corresponding to each frequency band (tree node) were integrated together, meaning that the energy for each node was computed to obtain 510 integration coefficients. This means that, by this process, 510 coefficients are obtained for each

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

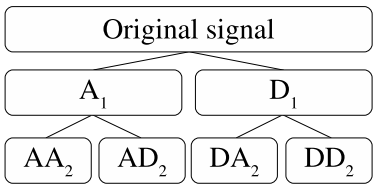

Fig. 5. Wavelet packets tree with two decomposition levels ("A" stands for *approximation* and "D" for *detail* coefficients).

signal frame. Then, for each signal (sound file) the average, deviation and kurtosis of each coefficient along the frames is computed. Then, for each sound file we obtain a total set of 1530 features from which a subset with the most relevant information is to be selected.

Different wavelet families have been compared in order to determine which one is the most convenient for this task. Based on the literature focused on speech analysis [34, 35, 43] and preliminary experiments that included different wavelet families, in this work we show the results obtained with the 4th- and 5th-order Coiflet and 8th-order Symlet families.

We propose the use of an EA for the selection of the optimal feature subset. The objective function should evaluate the representation suggested by a given chromosome (individual in the EA population), providing measures which are relevant for this particular problem. The candidate solutions represented by the individuals in the population of the EA are defined by binary chromosomes composed of 1530 genes, each of which corresponds to a particular wavelet-based feature. The target function evaluates the selected feature subset, providing a measure of classification performance. A classifier is used as objective function, so that the classification accuracy is obtained for each evaluated individual. This classifier is trained and the accuracy obtained is the fitness value for the corresponding chromosome.

2.6. Classifier

The Extreme Learning Machines (ELM) is a kind of artificial neural network with one hidden layer [28] and its main peculiarity respecting to classical models is the training algorithm. It does not need parameter tuning and the hidden neurons are randomly initialised. Consequently, the training time is significantly reduced with respect to other training methods that use complex optimisation techniques.

Formally, let be J hidden units with F inputs and P output units. The hidden layer output is given by

$$h_j = \Phi(\mathbf{v}_j^T \mathbf{x} + b_j), \tag{13}$$

with Φ as a non-linear activation function, $\mathbf{v}_j$ the input weights and b_j the bias for the j-th hidden unit. The hidden-layer output, also known as projected features, can be expressed as $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_N]^T$. Rewriting the equation in a matrix form, with $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_P]$, $\mathbf{w}_p \in \mathbb{R}^J$ and $p = 1, \ldots, P$ as the output layer weights, the ELM output is

$$\tilde{\mathbf{Y}} = \mathbf{HW}. \tag{14}$$

If the function Φ satisfy certain properties (infinitely differentiability and random hidden weights) it can be shown that for any pair of inputs $(\mathbf{X}, \mathbf{Y})$ exists a number $J < N$ such $||\tilde{\mathbf{Y}} - \mathbf{Y}|| < \epsilon$ for any small ϵ [28]. This means that the ELM can approximate the target $\mathbf{Y}$ as much as we want by adjusting only the number of hidden units and the output weights. The optimisation problem for $\mathbf{W}$ can be written as

$$\underset{\mathbf{W}}{\text{minimise}} \ ||\mathbf{HW} - \mathbf{Y}||_2, \tag{15}$$

which is a least square optimisation problem. The smallest norm solution is given by

$$\hat{\mathbf{W}} = \mathbf{H}^{\dagger}\mathbf{Y}, \tag{16}$$

where $\mathbf{H}^{\dagger}$ is the Moore-Penrose pseudo-inverse [10]. This solution for the optimisation problem is greatly fast comparing with the classical classifiers as SVM or back-propagation multi-layer perceptrons. More mathematical details of the ELM algorithm and several comparison with other neural nets can be seen in [27, 28].

3. Results and discussion

For the experiments, a set of audio recordings from the CRIED corpus was used. Since the examples composing the test set of the CRIED database are not labelled, only the train set consisting on 2838 instances was used in this work. Each of the instances in the train set is labelled as one of three categories: *Positive Mood* (2292), *Fussing* (368) or *Crying* (178). The experiments were carried out with a stratified cross-validation schemed in 10 folds, and the best results for different configurations of the ELM classifier and various feature types are presented. Since the data-set is not balanced, in order to evaluate the performance

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

Table 2
Summary of the best results (10-fold average).

Features	FV size	UAR[%]	ACC[%]
Baseline (MFCC & funct.)	531	62.15	79.84
MLS	110	65.88	85.73
MLSa	110	68.61	87.88
all MLS+MLSa	328	67.37	85.73
MLS+MLSa (Added)	230	68.76	87.74
MLS+MLSa (Combined)	230	68.94	86.82
4 NNAPP (512 - 2x)	256	57.75	84.53
8 NNAPP (2560 - 10x)	256	58.17	73.18
16 NNAPP (1280 - 5x)	256	60.67	79.92
32 NNAPP (768 - 3x)	256	63.55	84.99
64 NNAPP (1280 - 5x)	256	60.88	80.37
Symlet8	[263]	54.55	77.57
Symlet8 + MLS	≈ 463	62.33	83.07
Symlet8 + MLSa	≈ 391	60.43	81.41
Symlet8 + MLS + MLSa	≈ 591	62.33	84.21
Coiflet4	[267]	55.08	81.83
Coiflet4 + MLS	≈ 467	61.52	83.29
Coiflet4 + MLSa	≈ 395	60.68	83.93
Coiflet4 + MLS + MLSa	≈ 595	62.22	83.46
Coiflet5	[265]	55.98	81.97
Coiflet5 + MLS	≈ 465	60.67	85.03
Coiflet5 + MLSa	≈ 393	61.00	84.49
Coiflet5 + MLS + MLSa	≈ 593	61.34	84.81
ESCC	45	68.67	86.05
ESCC + MLS	155	68.30	85.16
ESCC + MLSa	155	**69.60**	**87.95**
ESCC + MLS + MLSa	265	69.04	87.91

appropriately the Unweighted Average Recall (UAR) [41] measure was considered, in addition to the classification accuracy.

Table 2 shows the results obtained in the evaluation of the different feature sets (FV stands for *feature vector*). The first row shows the Baseline results and the following five rows show the results for features MLS and MLSa, which were evaluated separately and combined together. In Table 2, "all MLS+MLSa" refers to the feature set composed of all the MLS and MLSa coefficients, without reducing dimensionality with F-Score. Also, the MLS and MLSa feature set were combined to apply F-Score for dimensionality reduction.

When reducing dimensionality with F-Score, in order to select the appropriate number of features to maintain, the classification performance is evaluated for incremental feature subsets containing the top ranked features. The subset of the top 10 features is evaluated first, then the top 20 and so on. Then, the subset that provides the best performance is kept. In this manner, it was determined that for both MLS and MLSa the best feature subset consists of the first 110 features in the rank. The MLS and MLSa were combined applying F-Score first to keep the 110 best features from each set (Added), and were also combined all together to apply F-Score keeping the 230 best features from the complete set (Combined). The last section of the table also shows MLS and MLSa combined with ESCC features.

Table 3
Results using NNAPP on frame level classification.

Features	FV size	UAR[%]	ACC[%]
4 NNAPP	256	55.71	71.46
8 NNAPP	256	58.12	73.25
16 NNAPP	256	60.53	74.97
32 NNAPP	256	62.65	76.00
64 NNAPP	256	**62.89**	**76.30**

The third block of Table 2 shows the results for the sparse representation. Here, $\mathbf{x} \in \mathbb{R}^{64\times 4}$ (sliding windows of 64 coefficients for the auditory spectrogram by 4 frames). The number of atoms in the dictionary is $M = 256$, which is the FV size for all the cases. The number of selected atoms for the reconstruction (NNAPP) were varied from 4 to 64. The ELM hidden layer size in the classifiers was varied from 256 (1x) up to 2560 (10x the feature vector size). The small numbers in brackets in the form (hidden - multiplier) detail the number of hidden units in the ELM classifier that obtained the best results for each NNAPP case. It can be noted that the best configuration in terms of UAR corresponds to 32 non-zero coefficients in the sparse representation with an ELM hidden layer size of 3 times the FV size.

The fourth, fifth and sixth blocks of Table 2 show the results obtained with the optimised wavelet features presented in Section 2.5.2. Each of these blocks present the results of exactly the same experiments though considering different wavelet families (Symlet8, Coiflet4 and Coiflet5). Since the EA based feature selection was repeated for each CV fold, and the number of target features is not fixed, an average of the number of selected features across folds is shown in the table (between brackets). As it can be seen in the table, all the optimised wavelet feature sets where evaluated alone and combined with MLS and MLSa coefficients. The corresponding dimensionality shown in the table for

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

each feature combination is adding the number of coefficients for MLS and MLSa to the average number of wavelet features. Here we considered the full sets of MLS and MLSa features consisting on 200 and 128 coefficients, respectively. This means that, for each fold, the actual number of features differs from the number indicated in the table, and this is the meaning for the symbol $\approx$. It is interesting to note that the average number of selected features are close for all the wavelet families. The results show that the average performance is quite similar for all the three wavelet families. Also, for all Coiflet4, Coiflet5 and Symlet8 the combination with both MLS and MLSa features allows to improve the performance. However, as it can be seen in the table, different combinations of MLS and MLSa alone provide much better performances. This may be due to the dimensionality when all the features are combined.

The last block in Table 2 shows the performances obtained with the proposed ESCC features as well as the combination of ESCC with MLS and MLSa. As explained in Section 2.4.3, the ESCC parameterisation consists on 45 features and, in this case, only the first 110 features were considered for both MLS and MLSa to be combined with ESCC.

It can be seen in the table that the MLS, MLSa and ESCC feature sets significantly outperform the Baseline in both UAR and accuracy (ACC). Moreover, different combinations of these feature sets are able to provide even better performance. Also, it is important to note that all of the representations proposed have lower dimensionality than the Baseline. For instance, the ESCC features provides an improvement of 6.52% of UAR with less than 10% of the attributes of the Baseline, showing that this representation is much more convenient for this task. The combination of MLS and MLSa also improves their individual performances when the F-Score measure is applied to keep the most discriminative attributes. Finally, the best result is provided by the combination of ESCC and MLSa, in both UAR and Accuracy, with a relatively small feature set. Also, the optimised wavelet features combined with MLS and MLSa, and the NNAPP features provided performances similar to the baseline with fewer coefficients. Furthermore, the 32-NNAPP was able to outperform the baseline.

These results are also relevant when compared with recent works on infant mood classification using the same dataset. In [29], for example, an UAR of 68.72% was obtained by a combination of static and dynamic neural network and a large set of features. This means

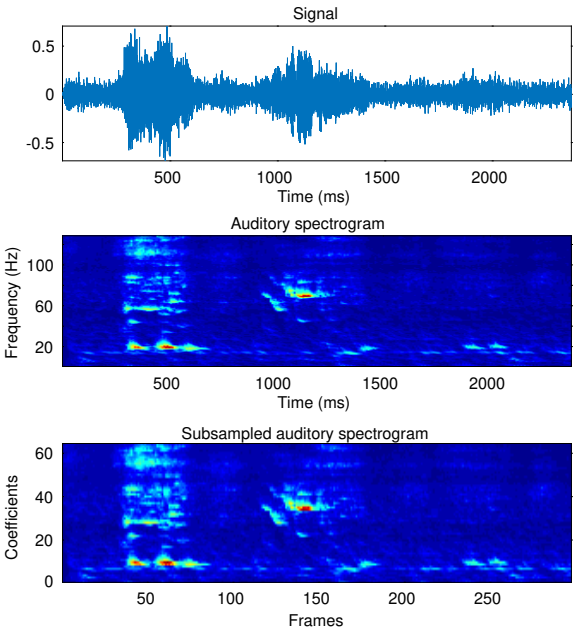

Fig. 6. Example of an audio signal of the database and the corresponding auditory spectrogram (original and 64 coefficients version).

that the combination of ESCC and MLSa features in our approach was able to improve this result using a simpler static classification scheme. Also a close result of 71.60% UAR was obtained by means of convolutional neural networks trained over spectrograms in [50]. However, the results can not be compared without taking into account that theirs was obtained using the test set of CRIED corpus, while we only used the train set (as explained at the beginning of this section). Which means that we used a smaller number of instances for training our models and, therefore, it is expected to obtain lower performance.

For illustration purposes, Figure 6 shows an audio signal of a neutral/positive mood vocalisation along with the auditory spectrogram. The spectrograms span from 0 to 8 kHz, with a posterior sampling in 64 coefficients. From a subset of these spectrograms the dictionary is computed by means of the NN-K-SVD algorithm using an atom width of 4 frames. Figure 7 shows an excerpt of the dictionary obtained (only 25 from 256 atoms are shown). It can be seen that the atoms learned different features appearing in the spectrograms, for example pure tones of different frequency (2nd column), start/stop of events (5th column) and background or vocalisation noises (first two atoms in 1st column). Finally, Figure 8 shows the 17 feature vectors **a** with 256 coefficients in columns, obtained

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

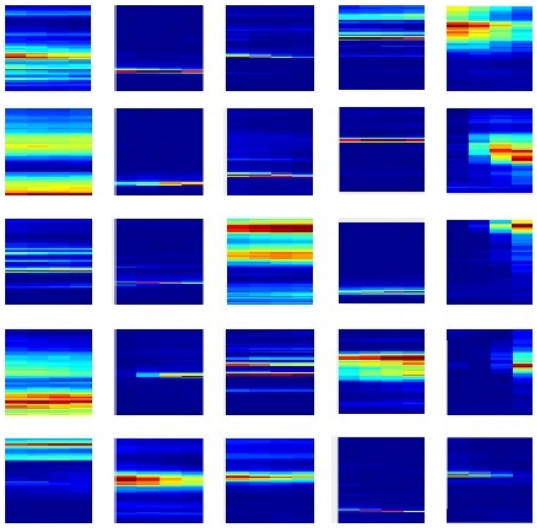

Fig. 7. Dictionary obtained from the auditory spectrograms.

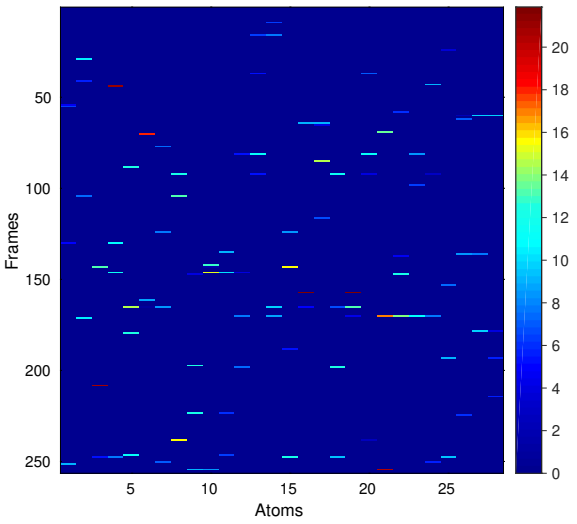

Fig. 8. Feature vectors (frame level classification) with the activations for the signal of Figure 6.

for the signal of Figure 6 using NNAPP=4 (the scale for the activation energy is given as reference).

Figure 9 shows the filter bank that was obtained by the optimisation process for the ESCC features. As it can be seen, the information on frequency band from 500Hz to 2500Hz, approximately, is enhanced with higher amplitudes in this filter bank. This corresponds to the frequency bands that show more inter class variance in the corpus (as seen in Figure 3). Also, at low frequencies (below 1000Hz) it shows higher resolution to capture the information related to the peaks in the mean log-spectrums of Figure 3. These remarks, together with the results obtained, show that the optimisation provided a filter bank that is much more appropriate for this task.

In order to consider and compare a different level of analysis, we performed the classification at the frame level (instead of averaging features for each sound file). The results using different configurations with NNAPP are shown in Table 3. Since the performance is similar to that shown in Table 2, these results show that no relevant information is lost when averaging features along time frames. Moreover, this suggests that the information contained in an isolated frame is useful for the classification and dynamic information might not be essential for this task.

In order to provide a qualitative analysis of the optimized wavelet decomposition, Figure 10 shows the relative relevance of each frequency band given by the wavelet packets tree. The relevance is given by the number of times that the corresponding node was selected by the EA, considering all the optimisation experiments performed for each data fold and wavelet family. The frequency resolution varies for each decomposition level, so that the number of bands in each level is 2^j, being j the decomposition level. Since the recordings were downsampled to 8kHz, the total available bandwith goes from 0 Hz to 4 Khz (similar to the filter bank in Figure 9). Brighter colors indicate frequency bands with higher relevance for this task. As it can be seen, the interpretation of this figure is not as straightforward as Figure 9. However, focusing on decomposition level 3, for instance, it can be seen that the first half of the frequency bandwith is given higher relevance. This agrees with the analysis of the optimised filter bank. Moreover, considering all the decomposition levels, one of the most relevant frequency bands is the one from 0 to 500Hz (at level 3), and this is also the band with higher filter resolution in the filter bank (Figure 9).

4. Conclusions and future work

In this work several feature sets were proposed to improve the performance in infant mood classification, which is a challenging and relevant problem to be tackled by the affective computing community.

The proposal relies on five different feature sets based on: the mean log-spectrum, an auditory spectrum, sparse dictionary representation based on auditory spectrum, wavelet coefficients optimisation and filter bank optimisation using evolutionary strategies.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

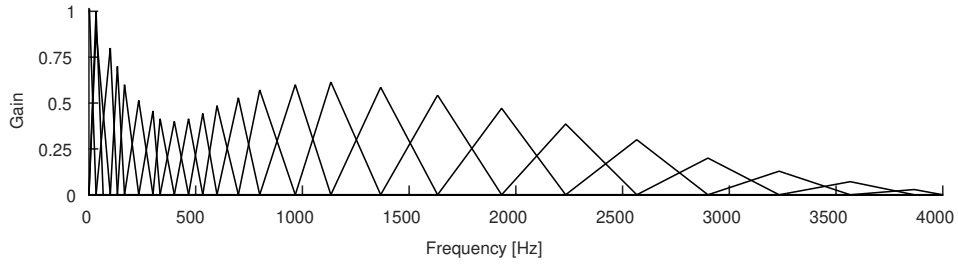

Fig. 9. Filter bank optimised for infant mood classification.

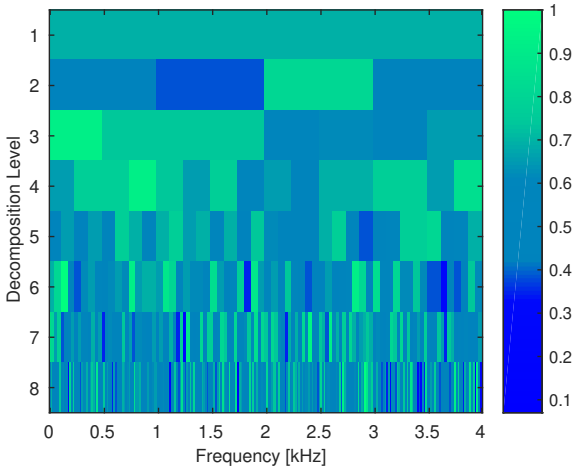

Fig. 10. Relevance of wavelet packets frequency bands.

The classification was carried out using a state-of the art neural network, the Extreme Learning Machines, in a cross-validation scheme.

The obtained performances in many cases outperformed the proposed baseline, showing significantly improved results with reduced sets of features. The results also showed that the proposed features are useful as improved representations for the classification of infant cry and mood. They also suggest that it is possible to improve the classical mel filter bank and to adapt wavelet based features for specific tasks. Also, the sparse dictionary representations based on auditory spectrograms have shown to be feasible for this task.

It is important to note that this study was limited to clean signals; however, the impact of noise on the performance of the different feature sets should be evaluated. Moreover, the effect of noise in the selection of wavelet features and in the shape of the evolved filter banks needs to be studied. Thus, further experiments will include signals under different types and levels of noise in order to evaluate and compare the robustness of the different feature sets. Then, the corpus used for evolving filter banks and selecting wavelet features will contain noisy signals. In this way we expect that, in both approaches, the optimisation would enhance those frequency bands which are less degraded by noise and can still provide useful information. In order to develop a robust classification system, it would also be interesting to increase the size of the corpus by including examples recorded in different conditions, regarding environmental situations like reverberation and diverse types of recording devices. It would also be interesting to evaluate the performance of these feature sets considering other types of mood and cries. All of these situations would probably require to repeat the optimisation procedure for both filter banks and wavelet features, in order to obtain the best possible performance.

Also, regarding the optimisation of filter banks, other parameters such as the number of filters, filter shape and filter bandwidth could be optimised in future developments. Another important matter in which future research will focus is to explore other ways for combining the different feature sets in order to better exploit their properties.

Acknowledgements

The authors wish to thank the support of the *Agencia Nacional de Promoción Científica y Tecnológica* (with PICT 2015-0977), the *Universidad Nacional de Litoral* (with CAI+D 50020150100055LI, CAI+D 50020150100059LI, CAI+D 50020150100042LI), and the *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) from Argentina.

References

[1] L. Abou-Abbas, C. Tadj, and H. A. Fersaie. A fully automated approach for baby cry signal segmentation and boundary detection of expiratory and inspiratory episodes. *The Journal of the Acoustical Society of America*, 142(3):1318–1331, 2017.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

[2] R. K. Aggarwal and M. Dave. Filterbank optimization for robust ASR using GA and PSO. *International Journal of Speech Technology*, 15(2):191–201, Jun 2012.

[3] Aharon, M and Elad, M. and Bruckstein, A. K-SVD and its non-negative variant for dictionary design. In *Proceedings of the SPIE conference wavelets*, volume 5914, 2005.

[4] K. S. Ahmad, A. S. Thosar, J. H. Nirmal, and V. S. Pande. A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network. In *2015 Eighth International Conference on Advances in Pattern Recognition (ICAPR)*, pages 1–6, Jan 2015.

[5] E. M. Albornoz, D. H. Milone, and H. L. Rufiner. Spoken emotion recognition using hierarchical classifiers. *Computer Speech and Language*, 25(3):556–570, 2011.

[6] E. M. Albornoz, D. H. Milone, and H. L. Rufiner. Feature extraction based on bio-inspired model for robust emotion recognition. *Soft Computing*, 21(17):5145–5158, Sep 2017.

[7] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43(2):155–177, Feb 2015.

[8] V. Arora, P. Sood, and K. U. Keshari. A stacked sparse autoencoder based architecture for Punjabi and English spoken language classification using MFCC features. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 269–272, March 2016.

[9] H. Barlow. Redundancy reduction revisited. *Network: Computation in Neural Systems*, 12:241–253, 2001.

[10] A. Ben-Israel and T. N. E. Greville. *Generalized inverses: theory and applications*. Springer, 2 edition, 2001.

[11] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun. Short-term emotion assessment in a recall paradigm. *International Journal of Human-Computer Studies*, 67(8):607–627, 2009.

[12] Y.-W. Chen and C.-J. Lin. *Combining SVMs with Various Feature Selection Strategies*, pages 315–324. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[13] T. Chi, P. Ru, and S. A. Shamma. Multiresolution spectrotemporal analysis of complex sounds. *Journal of the Acoustical Society of America*, 118(2):887–906, 2005.

[14] R. Coifman and M. V. Wickerhauser. Entropy-based algorithms for best basis selection. *IEEE Transactions on Information Theory*, 38(2):713–718, 1992.

[15] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

[16] D. Klein, P. Konig, K. Kording. Sparse spectrotemporal coding of sounds. *EURASIP Journal on Applied Signal Processing*, 2003(7):659–667, 2003.

[17] B. Delgutte. Physiological models for basic auditory percepts. In H.H. Hawkins, T.A. McMullen, A.N Popper, R.R. Fay, editor, *Auditory Computation*. Springer, New York, 1996.

[18] J. Deller, J. Proakis, and J. Hansen. *Discrete-time processing of speech signals*. Macmillan Pub. Co., 1993.

[19] J. E. Drummond, M. L. McBride, and C. F. Wiebe. The development of mothers' understanding of infant crying. *Clinical Nursing Research*, 2(4):396–410, 1993. PMID: 8220195.

[20] E. Oja, A. Hyvärinen. Independent component analysis: a tutorial. *Neural Networks*, 13(4-5), 2000.

[21] F. Eyben. *Real-time Speech and Music Classification by Large Audio Feature Space Extraction*. Springer Theses. Springer International Publishing, 2015.

[22] F. Theunissen, K. Sen, A. Doupe. Spectro-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *Journal of Neuroscience*, 20:2315–2331, 2000.

[23] A. R. Ferreira da Silva. Approximations with evolutionary pursuit. *Signal Processing*, 83(3):465–481, 2003.

[24] J. O. Garcia and C. A. R. Garcia. Mel–frequency cepstrum coefficients extraction from infant cry for classification of normal and pathological cry with feed-forward neural networks. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, volume 4, pages 3140–3145, July 2003.

[25] D. Giakoumis, D. Tzovaras, and G. Hassapis. Subject-dependent biosignal features for increased accuracy in psychological stress detection. *International Journal of Human-Computer Studies*, 71(4):425–439, 2013.

[26] L. Gu and K. Rose. Perceptual harmonic cepstral coefficients for speech recognition in noisy environment. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 1, pages 125–128 vol.1, 2001.

[27] G. Huang, G.-B. Huang, S. Song, and K. You. Trends in extreme learning machines: A review. *Neural Networks*, 61:32 – 48, 2015.

[28] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[29] M. Huckvale. Neural network architecture that combines temporal and summative features for infant cry classification in the interspeech 2018 computational paralinguistics challenge. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 137–141. International Speech Communication Association (ISCA), 2018.

[30] J. Hung. Optimization of filter-bank to improve the extraction of MFCC features in speech recognition. In *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, pages 675–678, Oct. 2004.

[31] A. Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum-likelihood estimation. Technical report, Helsinki University of Technology, 1998.

[32] S. Lee, S. Fang, J. Hung, and L. Lee. Improved MFCC feature extraction by PCA–optimized filter–bank for speech recognition. In *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, pages 49–52, 2001.

[33] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju. Speech based human emotion recognition using MFCC. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 2257–2260, March 2017.

[34] Y. Long, L. Gang, and G. Jun. Selection of the best wavelet base for speech signal. In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing*, pages 218–221, Oct 2004.

[35] S. Mallat. *A Wavelet Tour of signal Processing*. Academic Press, 3º edition, 2008.

[36] P. B. Marschik, F. B. Pokorny, R. Peharz, D. Zhang, J. O'Muircheartaigh, H. Roeyers, S. Bölte, A. J. Spittle, B. Urlesberger, B. Schuller, et al. A novel way to measure and predict development: a heuristic approach to facilitate the early

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.

detection of neurodevelopmental disorders. *Current Neurology and Neuroscience Reports*, 17(5):43, 2017.

[37] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[38] A. L. Oliveira, P. L. Braga, R. M. Lima, and M. L. Cornélio. GA-based method for feature selection and parameters optimization for machine learning regression applied to software effort estimation. *Information and Software Technology*, 52(11):1155 – 1166, 2010.

[39] S. Paul and S. Das. Simultaneous feature selection and weighting - an evolutionary multi-objective optimization approach. *Pattern Recognition Letters*, in press, 2015.

[40] O. F. Reyes-Galaviz and C. A. Reyes-Garcia. A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks. In *SPECOM-2004, 9th Conference Speech and Computer*, 2004.

[41] A. Rosenberg. Classifying skewed data: Importance weighting to optimize average recall. In *INTERSPEECH 2012*, Portland, USA, 2012.

[42] R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *IEEE Transactions on Signal Processing*, 58(3):1553–1564, March 2010.

[43] H. Rufiner and J. Goddard Close. A method of wavelet selection in phoneme recognition. In *Circuits and Systems, 1997. Proceedings of the 40th Midwest Symposium on*, volume 2, pages 889–891, 1997.

[44] N. Saito and R. Coifman. Local discriminant bases and their applications. *Journal of Mathematical Imaging and Vision*, 5(4):337–358, 1995.

[45] B. Schuller, S. Steidl, A. Batliner, Baumeister, et al. The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & self-assessed affect, crying & heart beats. In *Computational Paralinguistics Challenge, Interspeech 2018*, 2018.

[46] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski. The INTERSPEECH 2011 Speaker State Challenge. *Proc. Interspeech, ISCA*, pages 3201–3204, Aug. 2011.

[47] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan. An auditory-based feature for robust speech recognition. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4625–4628, April 2009.

[48] M. Skowronski and J. Harris. Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition. *The Journal of the Acoustical Society of America*, 116(3):1774–1780, Sept 2004.

[49] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, March 2016.

[50] M. A. T. Turan and E. Erzin. Monitoring infant's emotional cry in domestic environments using the capsule network architecture. *Proc. Interspeech 2018*, pages 132–136, 2018.

[51] P. Upadhyaya, O. Farooq, M. R. Abidi, and Y. V. Varshney. Continuous Hindi speech recognition model based on Kaldi ASR toolkit. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 786–789, March 2017.

[52] K. Veer and T. Sharma. A novel feature extraction for robust EMG pattern recognition. *Journal of Medical Engineering & Technology*, 40(4):149–154, 2016.

[53] L. D. Vignolo, D. H. Milone, and H. L. Rufiner. Genetic wavelet packets for speech recognition. *Expert Systems with Applications*, 40(6):2350–2359, 2013.

[54] L. D. Vignolo, D. H. Milone, and J. Scharcanski. Feature selection for face recognition based on multi-objective evolutionary wrappers. *Expert Systems with Applications*, 40(13):5077–5084, 2013.

[55] L. D. Vignolo, H. L. Rufiner, D. H. Milone, and J. C. Goddard. Evolutionary Cepstral Coefficients. *Applied Soft Computing*, 11(4):3419–3428, 2011.

[56] L. D. Vignolo, H. L. Rufiner, D. H. Milone, and J. C. Goddard. Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classification. *EURASIP Journal on Advances in Signal Proc.*, 2011:8:1–8:14, 2011.

[57] D. Wang, D. Miao, and C. Xie. Best basis-based wavelet packet entropy feature extraction and hierarchical eeg classification for epileptic detection. *Expert Systems with Applications*, 38(11):14314 – 14320, 2011.

[58] S. M. Woolley, T. E. Fremouw, A. Hsu, and F. E. Theunissen. Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature neuroscience*, 8(10):1371–1379, 2005.

[59] Z. Wu and Z. Cao. Improved MFCC-Based Feature for Robust Speaker Identification. *Tsinghua Science & Technology*, 10(2):158–161, 2005.

[60] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory*, 38(2):824–839, march 1992.

[61] L. Zão, D. Cavalcante, and R. Coelho. Time-frequency feature and AMS-GMM mask for acoustic emotion classification. *Signal Processing Letters, IEEE*, PP(99):1–1, 2014.

[62] A. Zabidi, W. Mansor, L. Y. Khuan, R. Sahak, and F. Y. A. Rahman. Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. In *2009 5th International Colloquium on Signal Processing Its Applications*, pages 204–208, March 2009.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
L. D. Vignolo, E. M. Albornoz & C. E. Martínez; "Exploring feature extraction methods for infant mood classification"
AI Communications, Vol. 3, No. 32, pp. 191-206, 2019.