Feature set optimisation for infant cry classification

L. D. Vignolo[1,2], E. M. Albornoz[1,2], and C. E. Martínez[1,3]

[1] Research Institute for Signals, Systems and Computational Intelligence (sinc(i))
Facultad de Ingeniería y Cs. Hídricas, Universidad Nacional del Litoral
CC217, Ciudad Universitaria, Paraje El Pozo, S3000, Santa Fe, Argentina
[2] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina
[3] Laboratorio de Cibernética, Facultad de Ingeniería,
Universidad Nacional de Entre Ríos
{ldvignolo, emalbornoz, cmartinez}@sinc.unl.edu.ar

Abstract. This work deals with the development of features for the automatic classification of infant cry, considering three categories: neutral, fussing and crying vocalisations. Mel-frequency cepstral coefficients, together with standard functional obtained from these, have long been the most widely used features for all kind of speech-related tasks, including infant cry classification. However, recent works have introduced alternative filter banks leading to performance improvements and increased robustness. In this work, the optimisation of a filter bank is proposed for feature extraction and two other spectrum-based feature sets are compared. The first set of features is obtained through the optimisation of filter banks, by means of an evolutionary algorithm, in order to find a more suitable speech representation for the infant cry classification. Moreover, the classification performance of the optimised representation combined with other spectral features based on the mean log-spectrum and auditory spectrum is evaluated. The results show that these feature sets are able to improve the performance for the cry classification task.

Index Terms: evolutionary algorithms, features optimization, crying classification

1 Introduction

Crying is an important communication tool for infants to express their emotional states and psychological needs [10]. 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 [10]. This motivates the work on the development of automatic tools for the analysis and recognition of infant cry applicable to real life.

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 are the most widespread features for any kind of sound signals [9]. Since their use is not limited to voice signals [27], as in speaker identification [3], emotional state recognition [17], or spoken language classification [7]. These features have also been used for tasks involving other sound signals such as music information retrieval [26] and the detection of acoustic events [33]. The MFCC features have also been used for the recognition of pathologies in recently born babies through their crying [21], for the analysis of infant cry with hypothyroidism [37] and for classification of normal and pathological cry [12]. 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. However, since the physiology of human perception is not yet fully understood, the parameters for the optimal filter bank are not known. Moreover, what is the relevant information contained in a signal spectrum depends on the application. Thus, it is doubtful that only one filter bank would be able to enhance the information that is relevant for any particular task. This has motivated the development of many approaches for tuning the filter bank in order to obtain better representations [2, 15, 16]. The use of a weighting function based on the harmonic structure was also proposed for improving the robustness of MFCC [13]. Similarly, other tuning to the parameters of the mel filter bank have been introduced [34, 36]. Although, to our knowledge, an evolutionary strategy for the optimisation of a filter bank for cry recognition has not yet been proposed.

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 [28, 19]. That is the case in [25], 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 [20, 30], like the optimisation of wavelet decompositions for speech recognition [29]. Also, many other strategies for developing optimised representation for speech related tasks have been presented [31, 32]. Evolutionary approaches have also shown success for the development of new features for stressed speech classification [6]. Although, the evolutionary optimisation of representations for the cry recognition task has not been explored.

This work tackles the automatic classification of crying vocalisations to allow automatic mood monitoring of babies for clinical or home applications [24]. Particularly, an approach based on an evolutionary algorithm (EA) for the optimisation of a filter bank for feature extraction is presented. The approach relies on an EA and introduces a scheme for parameter encoding based on spline interpolation, with the goal of finding an optimised filter bank which takes part in the extraction of cepstral features. In this proposal the EA is designed to evolve a filter bank that is part of the process for computing cepstral features, using a classifier to assess the fitness in the evaluation of the evolved individuals. This approach provides an alternative representation to improve the performance of cry recognition.

In this work, the use of a set of features based on a bio-inspired model is also proposed. These features, which were first introduced for emotion recognition [5], are based on an auditory model to mimic the human perception [35]. Since these features have not yet been used for cry recognition, it is interesting to inquire if the properties provided by the auditory model are useful for this purpose.

2 Materials and methods

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 [24]. 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) [24].

The database is split into training and test partitions. The utterances were classified into the following three categories: (i) neutral/positive mood vocalisations, (ii) fussing vocalisations, and (iii) 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 [18]. In the experiments only audio recordings were considered and, since the labels for the instances of the test partition are not available, cross validation 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 MFCCs [9] was considered as a baseline. 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 Evolutionary filter bank optimisation

In order to analyse the appropriateness of the mel filter bank 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 1, 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 1. These plots present peaks at high

Table 1. Functionals applied to MFCCs [11, 23].

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

frequency bands showing different relative energy and shape, which could be useful for classification. Since the mel filter bank (shown on top of Fig. 3) 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.

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 $(x = 1, y = 1)$. 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

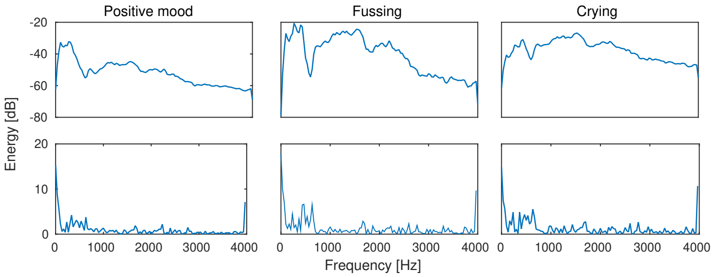

Fig. 1. Mean log-spectrums (top) and first-order difference of mean log-spectrums (bottom) for each of the three classes in the Cry Recognition In Early Development (CRIED) database.

each of the n_f filters, f_i^c, were obtained as

$$f_i^c = \frac{(y[i] - y_m) f_s}{y_M - y_m}, \tag{1}$$

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 $[0, 1]$.

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 2, where the gain of each filter was set according to the value given by spline II at the corresponding points.

2.3 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 2). 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

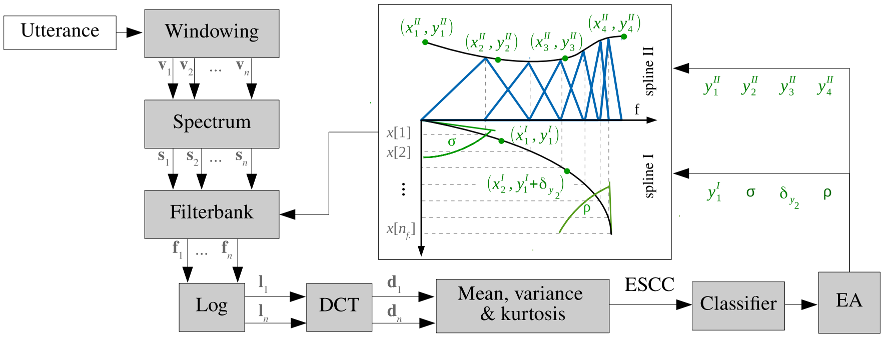

Fig. 2. Schematisation of the optimisation strategy. The output vectors of each block, $\mathbf{s}_i$, $\mathbf{f}_i$, $\mathbf{l}_i$ and $\mathbf{d}_i$, indicate that each window $\mathbf{v}_i$ is processed isolated and, finally, the mean and variance for each coefficient is computed from the $\mathbf{d}_i$ vectors in order to feed the classifier.

number of free parameters in the optimisation was reduced from 46 to 8 (4 parameters for each spline). The spline parameters were randomly initialized in the chromosomes using uniform distribution.

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 [31, 32]. 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 operator and the modifications were performed using a uniform random distribution.

2.4 Log-spectrum and auditory-spectrum based coefficients

A set of features obtained from the mean of the log-spectrum (MLS) was also considered. The MLS is defined as

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

where k corresponding to the frequency band, N is the total number of frames in the utterance, and $f(n,k)$ is the discrete Fourier transform of the signal in frame n. The spectrograms were computed using from non-overlapped Hamming windows of 25 ms. For 16kHz sampled signals, in this way 200 coefficients corresponding to equally spaced frequency bands are obtained. This processing was successfully applied for different speech related tasks [4].

Another set of features is used as well, which is based on the auditory spectrogram and the neurophysiological model proposed by Yang et al. [35]. This model consists in two stages, though only the first one is used here, which corresponds to the early auditory spectrogram. In this spectrogram the frequency bands are not uniformly distributed and 128 coefficients are thus obtained.

The mean of the auditory spectrogram (MLSa) is computed 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 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[4].

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 [8]. 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:

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

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. In the second approach all the MLS and MLSa features are ranked together by F-Score, in order to select the higher ranked features.

2.5 Classifier

Extreme Learning Machines (ELM) [14] are proposed to learn on the non-linear feature set. The primary implementation of ELM theory is a type of artificial neural network with one hidden layer. The main differences with classical models are in the training algorithm. The hidden units are randomly generated, thus the parameter tuning of this layer is avoided. As a direct consequence, the training time is reduced significantly compared with other training methods that have to use more complex optimisation techniques.

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

Table 2. Summary of the best results on training.

Features	FV size	UAR[%]	ACC[%]
Baseline (MFCC & functionals)	531	62.15	79.84
MLS	110	65.88	85.73
MLSa	110	68.61	87.88
ESCC	45	68.67	86.05
all MLS+MLSa	328	67.37	85.73
MLS+MLSa (Added)	230	68.76	87.74
MLS+MLSa (Combined)	230	68.94	86.82
ESCC + MLS	155	68.30	85.16
ESCC + MLSa	155	**69.60**	**87.95**
ESCC + MLS + MLSa	265	69.04	87.91

3 Results and discussion

Since the examples composing the test set of the CRIED database are not labelled, for the experiments 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 are presented. Since the dataset is not balanced, in order to evaluate the performance appropriately the Unweighed Average Recall (UAR) [22] measure was considered, in addition to the classification accuracy.

Table 2 shows the results obtained in the evaluation of the different feature sets. The described feature sets (MLS, MLSa and ESCC) 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). As the table shows MLS and MLS where also combined, together and separately, with the ESCC features.

As it can be seen in Table 2, 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 these representations have

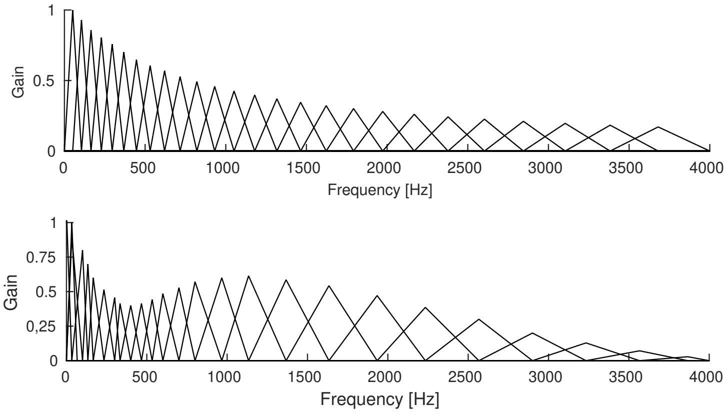

Fig. 3. Mel filter bank (top) and optimised filter bank (bottom).

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.

Figure 3 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 1). 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 1. These remarks, together with results obtained, show that the optimisation provided a filter bank that is much more appropriate for this task.

4 Conclusions

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

The proposal relies on three different feature sets: the first one based on the mean log-spectrum, a second feature set based on an auditory spectrum and the third one is optimised for this task by means of an evolutionary algorithm. The performance obtained through cross validation outperforms the baseline, showing significantly improved results with reduced sets of features.

The results show that the proposed features are useful as improved speech representations for cry recognition system, suggesting that there is further room for improvement over the classical mel filter bank for specific tasks.

It is important to note that this study was limited to clean signals, though it would be interesting to evaluate the impact of noise on the shape of the filter banks. Thus, further experiments will include noisy signals, as well as other types of cry and recording conditions. Also, other parameters regarding filter banks, such as the filter bandwidth could be also optimised in future work.

5 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. Abou-Abbas, L., Tadj, C., Fersaie, H.A.: 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). https://doi.org/10.1121/1.5001491, `https://doi.org/10.1121/1.5001491`
2. Aggarwal, R.K., Dave, M.: Filterbank optimization for robust ASR using GA and PSO. International Journal of Speech Technology **15**(2), 191–201 (Jun 2012). https://doi.org/10.1007/s10772-012-9133-9
3. Ahmad, K.S., Thosar, A.S., Nirmal, J.H., Pande, V.S.: 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). pp. 1–6 (Jan 2015). https://doi.org/10.1109/ICAPR.2015.7050669
4. Albornoz, E.M., Milone, D.H., Rufiner, H.L.: Spoken emotion recognition using hierarchical classifiers. Computer Speech and Language **25**(3), 556–570 (2011). https://doi.org/10.1016/j.csl.2010.10.001
5. Albornoz, E.M., Milone, D.H., Rufiner, H.L.: Feature extraction based on bio-inspired model for robust emotion recognition. Soft Computing **21**(17), 5145–5158 (Sep 2017). https://doi.org/10.1007/s00500-016-2110-5
6. Anagnostopoulos, C.N., Iliou, T., Giannoukos, I.: Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. Artificial Intelligence Review **43**(2), 155–177 (Feb 2015). https://doi.org/10.1007/s10462-012-9368-5
7. Arora, V., Sood, P., Keshari, K.U.: 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). pp. 269–272 (March 2016)
8. Chen, Y.W., Lin, C.J.: Combining SVMs with Various Feature Selection Strategies, pp. 315–324. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)

9. Davis, S.V., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics, Speech and Signal Processing **28**, 57–366 (1980)
10. Drummond, J.E., McBride, M.L., Wiebe, C.F.: The development of mothers' understanding of infant crying. Clinical Nursing Research **2**(4), 396–410 (1993). https://doi.org/10.1177/105477389300200403, pMID: 8220195
11. Eyben, F.: Real-time Speech and Music Classification by Large Audio Feature Space Extraction. Springer Theses, Springer International Publishing (2015), `https://books.google.com.ar/books?id=AFBECwAAQBAJ`
12. Garcia, J.O., Garcia, C.A.R.: 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. vol. 4, pp. 3140–3145 (July 2003). https://doi.org/10.1109/IJCNN.2003.1224074
13. Gu, L., Rose, K.: 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). vol. 1, pp. 125–128 vol.1 (2001). https://doi.org/10.1109/ICASSP.2001.940783
14. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541). vol. 2, pp. 985–990 vol.2 (July 2004). https://doi.org/10.1109/IJCNN.2004.1380068
15. Hung, J.: 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. pp. 675–678 (Oct 2004)
16. Lee, S., Fang, S., Hung, J., Lee, L.: Improved MFCC feature extraction by PCA–optimized filter–bank for speech recognition. In: Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on. pp. 49–52 (2001). https://doi.org/10.1109/ASRU.2001.1034586
17. Likitha, M.S., Gupta, S.R.R., Hasitha, K., Raju, A.U.: Speech based human emotion recognition using MFCC. In: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). pp. 2257–2260 (March 2017). https://doi.org/10.1109/WiSPNET.2017.8300161
18. Marschik, P.B., Pokorny, F.B., Peharz, R., Zhang, D., O'Muircheartaigh, J., Roeyers, H., Bölte, S., Spittle, A.J., Urlesberger, B., Schuller, B., et al.: A novel way to measure and predict development: a heuristic approach to facilitate the early detection of neurodevelopmental disorders. Current Neurology and Neuroscience Reports **17**(5), 43 (2017)
19. Oliveira, A.L., Braga, P.L., Lima, R.M., Cornlio, M.L.: 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). https://doi.org/https://doi.org/10.1016/j.infsof.2010.05.009
20. Paul, S., Das, S.: Simultaneous feature selection and weighting - an evolutionary multi-objective optimization approach. Pattern Recognition Letters **in press** (2015). https://doi.org/10.1016/j.patrec.2015.07.007
21. Reyes-Galaviz, O.F., Reyes-Garcia, C.A.: 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)
22. Rosenberg, A.: Classifying skewed data: Importance weighting to optimize average recall. In: INTERSPEECH 2012. Portland, USA (2012)

23. Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J.: The INTERSPEECH 2011 Speaker State Challenge. Proc. Interspeech, ISCA pp. 3201–3204 (Aug 2011)
24. Schuller, B., Steidl, S., Batliner, A., Baumeister, et al.: The interspeech 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In: Computational Paralinguistics Challenge, Interspeech 2018 (2018)
25. Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M.A., Schuller, B., Zafeiriou, S.: 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). pp. 5200–5204 (March 2016). https://doi.org/10.1109/ICASSP.2016.7472669
26. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing **10**(5), 293–302 (Jul 2002). https://doi.org/10.1109/TSA.2002.800560
27. Upadhyaya, P., Farooq, O., Abidi, M.R., Varshney, Y.V.: Continuous Hindi speech recognition model based on Kaldi ASR toolkit. In: 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET). pp. 786–789 (March 2017). https://doi.org/10.1109/WiSPNET.2017.8299868
28. Veer, K., Sharma, T.: A novel feature extraction for robust EMG pattern recognition. Journal of Medical Engineering & Technology **40**(4), 149–154 (2016). https://doi.org/10.3109/03091902.2016.1153739
29. Vignolo, L.D., Milone, D.H., Rufiner, H.L.: Genetic wavelet packets for speech recognition. Expert Systems with Applications **40**(6), 2350–2359 (2013). https://doi.org/10.1016/j.eswa.2012.10.050
30. Vignolo, L.D., Milone, D.H., Scharcanski, J.: Feature selection for face recognition based on multi-objective evolutionary wrappers. Expert Systems with Applications **40**(13), 5077–5084 (2013). https://doi.org/10.1016/j.eswa.2013.03.032
31. Vignolo, L.D., Rufiner, H.L., Milone, D.H., Goddard, J.C.: Evolutionary Cepstral Coefficients. Applied Soft Computing **11**(4), 3419–3428 (2011). https://doi.org/10.1016/j.asoc.2011.01.012
32. Vignolo, L.D., Rufiner, H.L., Milone, D.H., Goddard, J.C.: Evolutionary Splines for Cepstral Filterbank Optimization in Phoneme Classification. EURASIP Journal on Advances in Signal Proc. **2011**, 8:1–8:14 (2011)
33. Vozáriková, E., Juhár, J., Čižmár, A.: Acoustic Events Detection Using MFCC and MPEG-7 Descriptors. In: Dziech, A., Czyżewski, A. (eds.) Multimedia Communications, Services and Security. pp. 191–197. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
34. Wu, Z., Cao, Z.: Improved MFCC-Based Feature for Robust Speaker Identification. Tsinghua Science & Technology **10**(2), 158–161 (2005)
35. Yang, X., Wang, K., Shamma, S.A.: Auditory representations of acoustic signals. IEEE Transactions on Information Theory **38**(2), 824–839 (march 1992)
36. Zão, L., Cavalcante, D., Coelho, R.: Time-frequency feature and AMS-GMM mask for acoustic emotion classification. Signal Processing Letters, IEEE **PP**(99), 1–1 (2014). https://doi.org/10.1109/LSP.2014.2311435
37. Zabidi, A., Mansor, W., Khuan, L.Y., Sahak, R., Rahman, F.Y.A.: Mel-frequency cepstrum coefficient analysis of infant cry with hypothyroidism. In: 2009 5th International Colloquium on Signal Processing Its Applications. pp. 204–208 (March 2009). https://doi.org/10.1109/CSPA.2009.5069217