Testing the Universal Baby Language Hypothesis - Automatic Infant Speech Recognition with CNNs

Eduard Franti[1,2], Ioan Ispas[1,3], and Monica Dascalu[1,4]
[1] Research Institute for Artificial Intelligence, Bucharest, Romania, edif33@yahoo.com
[2] Simulation, Modelling and Computer-Aided Design Laboratory, IMT Bucharest, Romania, edif33@yahoo.com
[3] Applied Informatics Department, "Petru Maior" University, Targu Mures, Romania, ioanispas.john@yahoo.com
[4] Electronic Devices and Architectures Department, Politehnica University of Bucharest, Bucharest, Romania, monica.dascalu@upb.ro

***Abstract* — This paper presents an application of convolutional neural networks (CNN) for the recognition of the so-called "Dunstan baby language" that consists of five "words" or phonemes used by babies of age under 3 months to communicate their needs before they start crying. The model was derived from a CNN architecture which was successfully applied by the authors for voice-based emotion detection. The input of the neural network is the spectrogram obtained from the audio records of babies' voices and is processed as a two-dimensional image. The architecture was trained for a set of 250 small duration recordings and was tested for other 65 recordings with a recognition rate of 89%. The length of all audio files is less than 1 second; the recordings were extracted from certified Dunstan language recordings. The most important original contribution of the paper is the recognition of the actual "baby words" (and not the baby cry as was done before). This architecture offers an efficient tool for the verification of the "universal baby language" hypothesis, according to which the language of infants does not depend on culture, family, etc.**

***Keywords* — convolutional neural network; infant speech/crying classification; Dunstan baby language; universal baby language hypothesis**

I. INTRODUCTION

The accurate and fast interpretation of the babies' cries is important from the medical perspective, but also there is a huge interest to develop commercial devices and software applications for the families and caretakers of the babies [1]. According to Dustan's theory, before crying, the babies try to communicate their needs using a special „language" that consists of five "words" (or specific utterances) associated with five basic needs (like being hungry, sleepy etc.) [2]. The theory states that these five utterances are universal and innate.

A strong argument in favor of the "universal baby language" concept is the close relation of the mechanisms of emission of specific sounds and the particular physical problem that these sounds express. In Dunstan's perspective, the "infant words" should be properly and fast "decoded" by nurses, parents or caretakers, as little babies won't wait long. If the special need they expressed using the "baby language" is not taken care immediately, they will soon start to cry, because they already learned that cries attract help.

The possible applications of the Dunstan baby language (or infant speech) in pediatric hospitals, nurseries, and homes, are strongly limited by the recognition of this language. The usual approach implies human interpretation, that requires specific, certified training. An automatic method for this problem is still not available, as research in this direction was focused on cries classification.

This paper presents a neural network solution for infant speech recognition. The network classifies the audio recordings with baby utterances in 5 categories, corresponding to the 5 baby words of Dunstan's baby language. The problem was approached from an emotion detection perspective. The network was trained and subsequently tested with a set of audio files classified by certified Dunstan baby interpreters. Its future purpose is, in general, the development of infant cry/speech recognition and classification software, and in particular, the exploration (possible validation) of the hypothesis that Dunstan's baby language is universal.

The paper is organized as follows: section II presents a brief review of the state of the art in baby cry processing (which several authors call "baby language recognition"). Section III gives more specific details about the Dunstan baby language, its interpretation and details on the audio database we have used. Section IV describes the CNN architecture that was chosen and programmed for this application. Section V contains the discussion on practical experiments: training the network, verification and classification performances, as well as a comparison with other methods reported in the scientific literature. The paper ends with conclusions and several ideas for future research.

II. AUTOMATIC CLASSIFICATION OF BABY CRIES AND "BABY TALK"

There is a general consent that the infant cry (as an audio signal) contains valuable information, but also that it is hard to classify. According to [3], in the cry of a baby is embedded information about the physical condition, emotional state and identity of a baby (like gender, weight, health issues, etc.), as presented in Figure 1.

Several results were already reported in the scientific literature in the domain of automatic baby cry processing. As

the audio signal of the baby cry contains enough information for the identification of the baby, an identification system was proposed in [4] using only audio records of baby cries.

Health applications of baby cries classification try to identify sickness, pain or distress. Already in the 1990's, the automatic determination of the level of distress was proposed, based on the decomposition of the cry signals in basic "phonemes" or units [5]. The basic classification pain/no pain cry was approached with several methods belonging to speech processing domain with reported performances around 70% ([6], [7]).

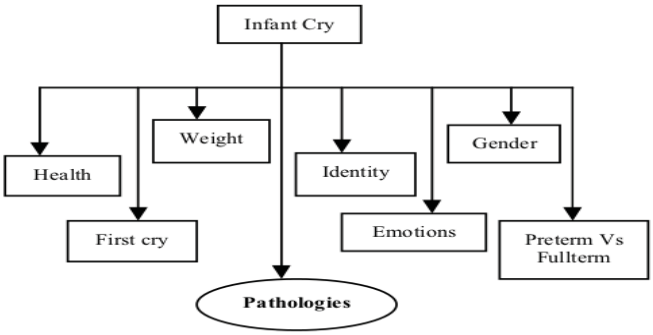

Fig. 1. Infant cry automatic classification and information extraction [3]

A neural-network based solution for identification of certain pathology is presented in [8]. The authors describe a system that classifies the cries of babies aged under 9 months in 3 classes: healthy babies, deaf babies and babies with respiratory problems (asphyxia).

Several researchers reported results in the direction of interpretation or "decoding" of the "baby talk" (i.e., of the baby cries) in relation to the Dunstan baby language. The background of these applications lies in the relation of the cry as an effect (or means of expression) of a physical (sometimes pathological) issue or distress. Babies cry for a reason, and that reason can be decoded through processing of the audio signal (the cry).

An automatic method of infant cry detection that is using speech features and the k-Nearest Neighbor Algorithm (KNN) is presented in [9]. LFCC (Linear Frequency Cepstrum Coefficients) proved to be more adequate than MFCC (Mel-frequency Cepstral Coefficients) for that particular task (with performances around 90% in recognition rate, for each of the five classes corresponding to the Dunstan baby language, compared to 79% with MFCC) [9].

Other authors have chosen MFCC and KNN classifier [10] and reported performances of 80% and 90% recognition rate for the 5 types of crying corresponding to the 5 words of Dunstan baby language. However, the authors have used recordings with children of age 1 day - 6 months, while Dunstan's theory refers to the age under 3 months.

Other methods proposed in scientific literature belong to language recognition area: the Gaussian Mixture Model - Universal Background Model and the i-vectors modeling methods", with an average accuracy of results of 70% [11].

The last three examples [9, 10, 11] refer to Dunstan' s baby language. All those methods classify the cries' signals in 5 classes, according to the 5 reasons for crying in Dunstan's theory. All methods use Dunstan's theory for a more objective, theoretically grounded, labeling of signals. In previous work, like [5], the interpretation of the cry recordings was done by an experimental group of parents. The cited papers do not question the validity of Dunstan's theory, nor do they recognize explicitly the five words of that language.

III. DUNSTAN BABY LANGUAGE: FACTS AND HYPOTHESIS

Priscilla Dunstan discovered in 2011 that little babies use a proto-language with five "words" to express their immediate needs. According to Dunstan, this proto-language (called since Dunstan baby language) is universal, meaning that all infants use the same (or very similar) five "words" in their first 3 months [2].

The five "words" of the infant "universal pro-language" were transliterated by Dunstan as: "Neh" = hungry; "Eh" = need to burp; "Oah (Owh)" = tired (sleepy); "Eairh (Eargghh)" = stomach cramp (lower gas); "Heh" = physical discomfort at skin kevel (feeling hot or wet, for example). According to Dunstan, infants first express a particular need with one of these phonemes. If their need is not taken care of, they will soon start to cry.

After 12 weeks of life, the infant will add some other phonemes and "words" to describe more complex needs and emotions. [2] Dustan's argumentation - apart from empiric evidence - is based on physiology. It is already proven that the physical reflexes related with particular needs produce specific sounds. For instance, "Eairh (Eargghh)" is coming "from the belly" by contracting the abdominal muscles when intestinal gases cause spasms.

Although Dunstan's theory is subject to criticism, it also began to be taken into consideration as research hypothesis. On the other hand, Dunstan's method of baby language interpretation had a huge success because of its practical applications [12].

The greatest difficulty of Dunstan's method of interpretation of infant communication is the human factor: specific training is required in order to understand the "baby words". In addition, the "verbalization" phase is short and therefore continuously attention and quick response are needed.

Automatic validation and recognition of the Dunstan's baby language was never done convincingly. As presented in the previous section, the methods proposed so far recognize the crying that follows the verbalization phase and therefore a quick response that prevent crying is not possible.

In our research we have used audio recordings extracted form Dunstan-certified video recordings available on the Internet, that contain a lot of examples of baby utterances, with their classification. Since the audio recording of the cry is not needed here, we were able to identify 315 such examples with certified interpretation as belonging to one of the 5 "words".

The methods for baby cries classification reported in scientific literature used fewer files (250 files in [10], 150 in [9], 50 in [5] and 40 in [11]).

IV. A DEEP LEARNING MODEL FOR DUNSTAN'S BABY LANGUAGE RECOGNITION

We have approached the classification of the audio signals corresponding to the 5 utterances of the Dunstan's baby language as an "affective computing" application. We have used a CNN architecture derived from an image processing CNN - that the authors have successfully used for voice-based emotion detection [13]. In [13], we have designed and trained the network for the detection of 6 basic emotions from voice. Figure 2 contains the block scheme of the architecture.

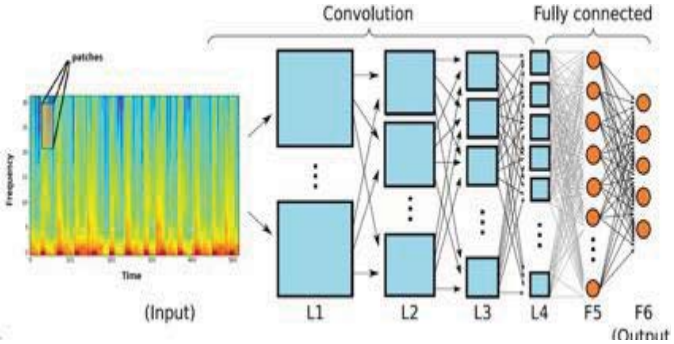

Fig 2. Generic representation of the CNN architecture [13]

The specificity of the CNN is the presence of pairs of convolutional and pooling layers. A convolutional layer has the purpose to extract the structured information with sub-matrices filters (strides) parsing on the two-dimensional input data. A pooling layer summarizes the output of the convolution matrix by aggregating the values of the stride sub-matrix into a single value. The CNN architecture includes also a number of dense (fully connected) layers, with the final (top) layer that contains the classifier [13].

TABLE 1. SUMMARY OF THE ARCHITECTURE (PYTHON RUNNING CODE)

```
>>> model.summary()
```

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 160, 500, 20)	520
max_pooling2d_1 (MaxPooling2	(None, 80, 250, 20)	0
flatten_1 (Flatten)	(None, 400000)	0
dense_1 (Dense)	(None, 128)	51200128
dense_2 (Dense)	(None, 6)	774

```
Total params: 51,201,422
Trainable params: 51,201,422
Non-trainable params: 0
```

For the implementation, we used Keras model-level library with the TensorFlow backend [14]. The architecture was programmed in Python based on the description of specific layers in Keras library [15]. After several experiments, the architecture selected for our application includes 20 convolutional filters of size 5x5, with ReLu activation, followed by a max-pooling and has 500 x 160 matrixes as input. The final stage consists of a flattening and a dense (hidden - fully connected) layer of 128 neurons, followed by the words classifier (Table 1 contains the summary of the architecture).

V. EXPERIMENTS AND RESULTS

A. Data set

The database consists of 315 audio files recorded at 16 kbps, with maximum 1 second length. Each file is the audio recording of a baby's utterance corresponding to one of the five "Dunstan words". The files are audio captures extracted by the authors, from the video materials released by Priscilla Dunstan, available on the Internet. The materials were recorded in studio conditions, without noises, echoes, etc. The labeling (classification) of the "words" was done by Dunstan herself or other Dunstan certified personnel. The audio captures do not require any other post-recording filtering. The data set contains an equal number of recordings for each "word": "neh", "eh", "owh", "eairh", "heh". 250 files (50 for each class) were used for the CNN training and 65 files were used for testing (13 for each class).

B. Data input (pre-processing)

For our application, because the length of the audio files is bellow one second, we have decided to use the full spectrogram as input data for the CNN, instead of other and features like MFCC or LPCC. The audio files were transformed to spectrogram using a PRAAT script. PRAAT is a free software [14] for speech analysis developed at University of Amsterdam.

The PRAAT parameters are: Window length = 0.005s; Maximum frequency = 5kHz; Time step = 0.002s and Frequency step = 20Hz. The spectrogram results as a 160 lines x 500 columns table of real values. So, the input of the convolutional network is a list of arrays of 160 x 500 values that can be simply looked like two-dimensional images. Figure 3 gives 5 examples of spectrograms, one for each of the 5 words (the images were obtained with Audacity audio software).

C. The CNN training and testing

Due to the complexity of the convolutional layers, the architecture contains over 5,200,000 trainable parameters.

The model was trained for 5 epochs using the 250 input files selected for the training and was tested with 65 the files selected for the testing. The files used for training/testing were randomly selected, but equally distributed within the 5 classes.

The final testing/evaluation error - on average - after multiple running is 11%, so the final recognition rate is 89%. The result is a good validation for the premise that there is no need for other preprocessing or audio feature extraction and we can use directly the spectrogram.

The CNN is fully able to learn and classify 315 baby words spectrogram images, with great amount of similarities in only 5 epochs. The explanation lies in the fact that the convolutional layer makes the feature extraction.

Although it is improper to compare out method with results reported in scientific literature (that were designed for

crying classification), the 89% accuracy can be appreciated as a promising result.

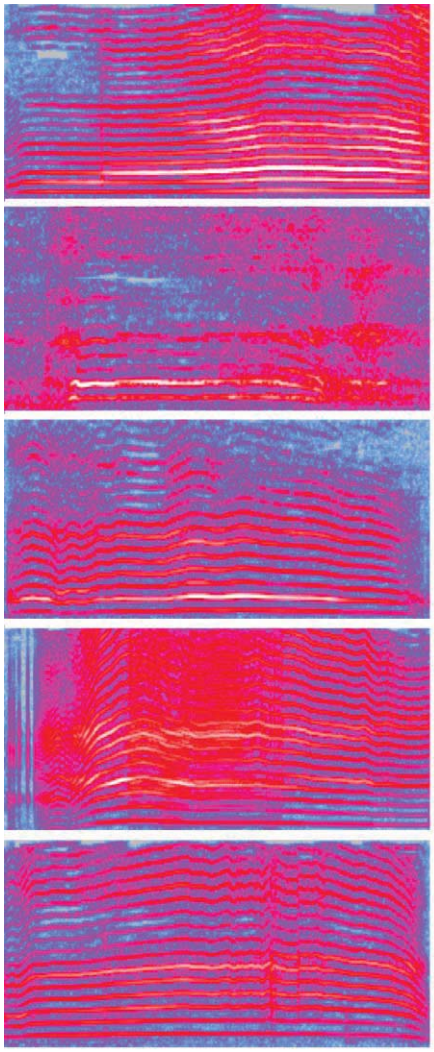

Fig 3. The spectrograms as images of the five "words": a) "Neh", b)"Heh", c) "Eh", d) "Eairh", e) "Oah"

VI. CONCLUSIONS

The paper presents an application of CNN for the recognition of the five words of the so-called "Dunstan baby language" corresponding to the following needs/emotions: hunger, burp, tired, cramp, discomfort. Our approach is different from other experiments on Dunstan baby language recognition as our aim was the classification of the "words" of that language (the utterances of the babies that precede the crying), while other results published in the scientific literature aim the recognition of the five types of cries.

Our method is completely oriented on the "words" of the Dunstan baby language. We have used 315 audio files (the files duration is below 1 second), that contain only the "words" and were labeled by certified Dunstan language interpreters. The input of the CNN is the spectrogram of the audio signal. The architecture was trained with 250 samples and tested on the rest of 65 files, with a recognition rate of 89%.

Another innovative perspective of our work is the fact that we offer a tool for the validation of the Dunstan's baby language hypothesis. Researchers that have used Dunstan's theory in their work on baby cries classification have considered that Dunstan's theory is 100% valid, while in the academic field of infantile psychology and pediatric medicine the theory is quite controversial.

In order to validate Dunstan's universal baby language hypothesis, the main direction of future research is to make experiments with recordings of babies from different cultural environments. If validated, our architecture will be implemented in a software application for real time detection and recognition of the infants' "universal language".

ACKNOWLEDGEMENTS

The work reported in this paper was partly supported by the European Project RoboCom++ FLAG-ERA JTC. The authors thank to the anonymous reviewers for their suggestions and comments that helped to improve this paper.

REFERENCES

[1] P. K. Kuhl, "Baby Talk", in Scientific American, November 2015, pp. 64-69, 2015.
[2] P. Dunstan, Calm the crying: Using the Dunstan baby language, London: Penguin Books Ltd., 2012.
[3] J. Saraswathy, et al. "Automatic classification of infant cry: A review", in Biomedical Engineering (ICoBE) Conference Proceedings, IEEE, pp. 543-548, 2012.
[4] A. Messaoud and C. Tadj. A cry-based babies identification system. In Proceedings of the 4th international conference on Image and signal processing, ICISP'10, pp. 192–199, 2010.
[5] Q. Xie, R.K. Ward, C.A. Laszlo, "Determining normal infants' level-ofdistress from cry sounds," Proc. of Canadian Conf. on Electrical and Computer Engineering, IEEE, pp. 1094-1096,1993.
[6] H.E. Baeck, M.N. Souza "A Bayesian classifier for baby's cry in pain and non-pain contexts," Proc. of the 25th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society, vol.3, pp.2944-2946, 2003.
[7] Y. Abdulaziz and S.M.S. Ahmad, "Infant cry recognition system: a comparison of system performance based on Mel frequency and linear prediction cepstral coefficients," Proceedings of Int. Conf. on Information Retrieval and Knowledge Management, Selangor, pp. 260-263, 2010.
[8] O.F. Reyes-Galaviz, C.A. Reyes-Garcia, "A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks", in 9th Conference Speech and Computer Proceedings, pp.1-6, 2004.
[9] V. Bhagatpatil, V.M. Sardar, "An Automatic Infant's Cry Detection Using Linear Frequency Cepstrum Coefficients (LFCC)", in International Journal of Scientific & Engineering Research, Vol. 5, Issue 5, pp.1379-1383, 2014.
[10] S. Bano, K. M. Ravikumar, "Decoding Baby Talk: Basic Approach for Normal Classification of Infant Cry Signal", International Conference on Current Trends in Advanced Computing (ICCTAC-2015) Proceedings, pp. 24-26, 2015.
[11] I.A. Bănică, H. Cucu, A. Buzo, D. Burileanu, C. Burileanu, "Automatic Methods for Infant Cry Classification", International Conference on Communications (COMM) Proceedings, IEEE, pp. 51-54, 2016.
[12] S. B. Lohre, "Attune with Baby: An Innovative Attunement Program for Parents and Families with Integrated Evaluation", Dissertations & Theses 350, 2017, retrieved March 2018 from aura.antioch.edu/etds/350
[13] E. Franti, I. Ispas, V. Dragomir, M. Dascalu, Z. Elteto; I. Stoica, "Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots" in Romanian Journal of Information Science and Technology, vol. 20, Issue 3, pp. 222-240, 2017.
[14] Keras: The Python Deep Learning library, Available at: https://keras.io.
[15] A. Gulli, S. Pal, Deep Learning with Keras, 2017, Packt Publishing
[16] P. Boersma, D. Weenink, Praat: Doing Phonetics by Computer (Version 6.0.35) [Computer Program]. Retrieved 2018 from http://www.praat.org/