Automatic Methods for Infant Cry Classification

Ioana-Alina Bănică, Horia Cucu*, Andi Buzo, Dragoş Burileanu and Corneliu Burileanu
Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Bucharest, Romania
*Corresponding author (E-mail: horia.cucu@upb.ro)

***Abstract*— Studies have shown that newborns are crying differently depending on their need: hunger, tiredness, discomfort, eructation, pain, and so on. Skilled persons such as pediatricians can distinguish between different types of newborn cries and consequently estimate the baby's need by using the sounds and the gestures produced by the baby. However, this is a real problem for the unskilled parents who would like to answer promptly to their baby needs. Recently, research work has been invested into developing automated methods to detect the baby's need using only a short audio recording of his/her live cry. In this paper we report the initial results we obtained by applying various methods that were successful in speaker and language recognition to the task of baby cry recognition. The results are promising: 70% accuracy on a dataset with 39 babies and 5 types of cry.**

Keywords - baby cry classification; infant cry recognition; GMM-UBM framework ; i-vectors framework.

I. INTRODUCTION

For newborns the only way of communication is by crying. In this way the newborn expresses his/her physical and emotional state and his/her needs. There are many reasons for a baby to cry, such as: hunger, tiredness, pain, discomfort, pain, colic, eructation and so on. For a parent it is very important to act as fast as possible to satisfy the baby's needs. Unfortunately, the parents are not always capable of understanding why their newborn is crying, so they become frustrated because they feel helpless in the face of the newborn's problem.

Skilled persons such as pediatricians can distinguish between different types of cries, finding some patterns in each type. However, it was noticed that the interpretations of cries can be subjective and the experience level of the listener is essential.

The newborn's cry can also be used to determine whether he/she suffers from pathological diseases, thus problems can be early discovered and this can be vital for the newborn. Cries which have an unusual sound, duration, intensity or height can show that the newborn is suffering from a specific disease [1]. For example the newborns who suffer of Down syndrome have a specific type of cry which allows pediatric specialists to detect this immediately.

A newborn cry is made of four types of sounds: one coming from the expiration phase, a brief pause, a sound coming from the inspiration phase, followed by another pause. Sound and pause durations and also the repetitions differ from child to child [2].

A newborn cry is a short-term stationary signal as is the speech. This signal is considered more stationary than the speech signal because newborns don't have full control of the vocal tract. Some new measurements have been defined to characterize and recognize the newborn cries, such as: latency, duration of expiratory component, the time between two cries, the value of the fundamental frequency, the dominant frequency, the amplitude of the dominant frequency, the global maximum amplitude [3].

Xie et al. [4] introduce a set of ten "cry phonemes" to characterize cry signals from normal infants to determinate the normal infants' level of distress from cry sounds. The ten "cry phonemes" were chosen to satisfy two criteria. The first is that the "cry phonemes" should constitute a basis for normal infant cry signals, thus together they cover most time-frequency patterns or variations commonly found in normal infant cries. The second criterion is that these phonemes should be detectable and distinguishable using computer signal analysis. They run an experiment in which 20 parents were instructed to repeatedly listen to the recordings of the 58 infant cries and they were asked to rate each cry into one of five levels of distress. It was observed that amongst the 10 "cry phonemes", dysphonation, hyperphonation, and inhalation show strong positive correlations with the parents' level of distress ratings, and flat and weak vibration show strong negative correlations. The dysphonation "cry phoneme" shows the most consistent positive correlation with the parents' ratings.

Baeck and Souza [5] built a Bayesian classifier for baby's cry in pain and non-pain context. For classification they used 25 cries recorded in a pain context and 25 cries in a non-pain context, separated in training and test groups in different folds. They obtain an accuracy around 75% that indicate a reasonable performance classification technique.

In [1] researchers developed a system for the processing of infant cry to recognize pathologies in newborns with neural networks. The system classifies three different kinds of cries, which come from normal, deaf and asphyxiating infants of ages from one day up to nine months old.

In this paper we present several baby cry classification experiments performed on a database of about 40 healthy babies, crying due to five physiological needs: hunger, discomfort, eructation, flatulence and tiredness. The database was created and labeled by our research group based on the Dunstan Baby Language (DBL) baby-cry classification video tutorial. For automatic classification, we employed two methods that were successful in speaker and language recognition: the GMM-UBM (Gaussian Mixture Model - Universal Background Model) and the i-vectors modeling methods.

The rest of the paper is structured as follows. Section II describes in detail the database used in the experiments. Section III presents briefly the well-known classification methods (GMM-UBM and i-vectors) used in Section IV for the experiments and finally Section V is dedicated to conclusions.

II. BABY CRY DATABASE

Collecting and more importantly labeling a database of baby cries is a very difficult and complex task. There are many issues

978-1-4673-8197-0/16/$31.00 ©2016 IEEE

regarding ethical and legal aspects that need to be taken into account and that hinder the development of such databases. First of all there is the problem of collecting baby cries in controlled environments, where doctors would know what is the need (are the needs) due to which the babies are crying. This can usually be achieved only in maternities. Creating and using a standard audio acquisition framework in a hospital is in itself a challenge.

Legal aspects refer to the need of obtaining the written consent of the parents regarding the acquisition and usage of the sounds produced by the babies, along with metadata regarding the babies.

Finally, finding the right medical personnel that is able to understand the reasons for which the babies are crying and label the corresponding audio files is again an obstacle.

The Speech and Dialogue research group [1] is one of the institutional partners in the SPLANN [2] research project, which aims to design and develop an automatic recognition system of the newborn cries. Within this project, "Sf. Pantelimon" Emergency Hospital in Bucharest is another institutional partner whose main role is to create an extensive database of infant cries labeled accordingly. The efforts implied by the database collection and labeling task and exhaustive details regarding the first version of the SPLANN database itself are provided by the coordinator of the project (Softwin Research) in [6].

While the activity described above is still in progress, we started the baby cries recognition research work on a smaller database, namely the Dunstan Baby Language database.

A. Dunstan Baby Language (DBL) Database

Dunstan Baby Pty Ltd [3] is a company dedicated to helping parents understand the needs of their crying babies. Following an 8-years research, they came up to the conclusion that "every baby, of every race, color and culture made the same 5 sounds before they cried out. And each cry meant baby needed something specifically". Priscilla Dunstan initiated this study after she figured out that she was able to find patterns in the cry of her own baby. She is also the presenter in a Dunstan Baby video tutorial which teaches parents to identify the five basic needs of newborns (hunger, discomfort, eructation, flatulence and tiredness) based on their cry. Each need was encoded with a set of phonemes, which sound similar with the newborns cry as follows: "EAIRH" – flatulence (C1), "EH" – eructation (C2), "HEH" – discomfort (C3), "NEH" – hunger (C4), "OWH" – tiredness (C5).

For starting the first experiments of automatic classification of baby cries, we used the audio-video footage presented by Priscilla Dunstan in the DBL tutorial. We extracted the audio from the DVD and selected the baby cries along with the information regarding the need given by the author. Using this methodology we were able to extract about 80 baby cries from 39 different babies, with a total duration of about 400 seconds. More details regarding each need are given in Table I. As noted from Table I, there are several babies for which we were able to extract more than one cry per need. Another observation is that the total number of cries (83) is much larger than the number of different babies (39): the same babies cry due to various needs. Table II lists the number of overlapping babies for the five needs in Table I.

[1] SpeeD Research Laboratory: speed.pub.ro
[2] SPLANN research project: softwinresearch.ro/index.php/ro/proiecte/splann
[3] Dunstan Baby Language: dunstanbaby.com

TABLE I. DETAILS FOR DUNSTAN BABY LANGUAGE DATABASE

Cry type	Code	# babies	# cries	Duration [s]
Flatulence (eairh)	C1	10	12	78
Eructation (eh)	C2	12	14	128
Discomfort (heh)	C3	12	13	43
Hunger (neh)	C4	23	25	60
Tiredness (owh)	C5	16	19	121
		Total	**83**	**430**

The statistics presented above were required for setting up the baby cry classification training/testing methodology. We divided the DBL database into two disjoint parts: a training part and an evaluation part. The division was made both at cry level and at subject level, such that all the babies used for evaluation are used for evaluation only (not also training). The evaluation part of the DBL database consists of all the cries available for 10 randomly selected babies.

Due to the fact that the DBL database is very small (in total only 430 seconds of baby cries), we used cross validation for all the experiments presented in Section IV. For this we divided the DBL database into training/evaluation sets 10 times, each time making sure that the evaluation part is different.

III. BABY CRY CLASSIFICATION METHODS

The baby cry classification task based solely on the audio signal is very similar to the task of phoneme classification, closed-set speaker identification and audio-based closed-set language identification. In these tasks the Mel Frequency Cepstral Coefficients (MFCCs) are for a long time the most preferred features thanks to their power of charactering human sounds and speech. Consequently, for the first experiments we performed for baby cry classification, we used the MFCC features to characterize the sounds produced by the babies.

Driven by the success of the GMM-UBM and i-vectors frameworks in modeling the speech of various speakers (in speaker identification [7, 8]), non-speech sounds such as music, jingles, noise and silence (in speaker diarization [9], for example) and various languages and dialects (in language and dialect identification [10]), we decided to start our investigation for baby cry classification with these two approaches.

A. The GMM-UBM Framework

For this method statistical models are built based on some vectors of features. The statistical model in this case is a Gaussian Mixture Model (GMM) which represents a probability distribution as a weighted sum of Gaussians. Each Gaussian is characterized by a mean and a variance.

TABLE II . OVERLAPPING BABIES PER TYPES OF CRY IN DBL DATABASE

# babies	C1	C2	C3	C4	C5
C1	10	3	2	4	3
C2		12	6	9	5
C3			12	8	7
C4				23	10
C5					16

In speaker recognition systems a universal background model (UBM) is a GMM representing general, person-independent speech. This notion can be adapted for newborn crying recognition systems: the UBM is here a model which represents general crying and is independent from the type of cry.

In the training stage, the UBM is estimated on the training database (comprising a large amount of audio representing general baby cry, not necessarily labeled with needs) using the Estimation Maximization (EM) algorithm. Going further, the UBM is adapted to the various types of baby cry (corresponding to the various needs) using need-specific audio. This adaptation is performed using Maximum A-Posteriori (MAP) algorithm. The output of this process is a set of GMMs, each of them adapted to a specific type of infant cry (i.e. hunger-GMM, tiredness-GMM, etc.).

In the evaluation stage the evaluation baby-cry files are scored against each GMM and the highest score is used to decide which the need was causing the baby to cry.

B. *The I-vectors Framework*

The i-vectors framework was also introduced for the first time in the context of speaker identification [11]. The method is also known as the total variability modeling matrix method. The main idea is to adapt the UBM to a set of characteristics based on eigenvoice adaptation technique [12], resulting a model specific for every type of cry.

Eigenvoice adaptation is based on the assumption that all important variability is contained in a matrix known as the total variability matrix (TVM) which is constructed in the training stage. The total variability matrix, unlike the eigenvoice space, contains on columns GMM supervectors (a concatenation of all mean vectors in a GMM), each created from one audio file in the training database.

In the training stage this method involves several steps:

- The UBM is estimated just like in the GMM-UBM case.
- The Total Variability Matrix (TVM) is constructed using the UBM and the training audio files.
- The TVM and need-specific cries are used to extract i-vectors for each type of cry.
- A Probabilistic Linear Discriminant Analysis (PLDA) Gaussian model is created for reducing the dimensionality of the i-vectors.

In the evaluation stage, the Cosine Scoring with WCCN normalization is used for classifying each cry recording to one of the cry classes.

IV. BABY CRY RECOGNITION EXPERIMENTS

A. *Experimental Setup*

The classification experiments were performed using the Microsoft Speaker Recognition (MSR) toolkit [13] and the LIUM toolkit [9].

For characterizing the type of cry we used 13 Mel Frequency Cepstral Coefficients (MFCC) extracted from overlapped windows (a similar setup as for speech and speaker recognition). The features were extracted using both MSR and LIUM toolkit. For the extraction of the MFCCs with MSR toolkit we used a Hamming window of size of 20 ms. For the extraction of parameters using the LIUM toolkit we used a Cosine window.

B. *Experimental Results*

1) *GMM-UBM classification using MSR toolkit*

To compensate the fact that the DBL database is very small, we used cross validation to validate our results. We performed the experiment ten times, each time choosing randomly a different set of 10 babies for evaluation and the rest of the babies for training.

As seen in Fig. 1 the best results were obtained by the model created with 8 Gaussian probability densities, achieving an average accuracy of 70%. The best accuracy (81.8%) was obtained for the evaluation set #5 for the model with 16 Gaussian densities.

We can conclude that increasing the number of densities per GMM above 32, no longer triggers increases in accuracy. This is because the model is already quite detailed, and adding more expressivity to the model causes over fitting.

2) *I-vectors classification using MSR toolkit*

In this experiment we used the same data set and the same MFCCs as in the previous experiment. The classification was made using the i-vectors framework.

For creating the UBM, training the PLDA and constructing the total variability matrix we used all the audio files available in the raw SPLANN database. The reason for doing this is because the DBL database was too small and, due to this, the over fitting phenomenon occurred: the classification process was not deterministic (successive runs of the same tests had substantially different results). The i-vectors specific for every type of cry were extracted using Dunstan Baby Language database.

We run the experiment ranging the number of Gaussians probability densities of the UBM from 4 to 256. The size of the TVM (100), the number of training iterations (5) and the size of LDA (5) remained constant.

Fig. 2 shows the results obtained for the first evaluation set. After this evaluation, we stopped the experiment due to unsatisfying results. The highest accuracy (47%) was obtained by the model with 8 Gaussian densities in the UBM (see Fig. 2).

3) *GMM-UBM classification using LIUM toolkit*

Although encouraging, the 70% classification accuracy obtained in experiment #1 was not considered satisfactory. Consequently, we reproduced the experiment using another toolkit that implements the training and scoring methods for the GMM-UBM framework, namely LIUM [9].

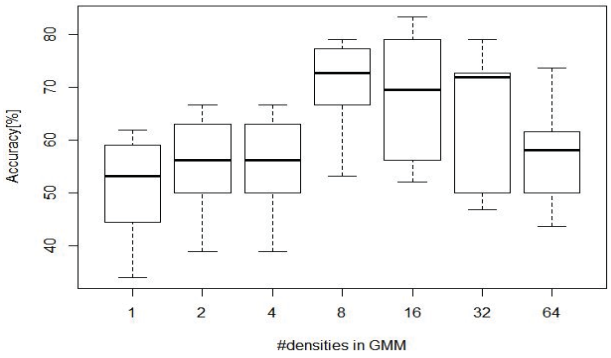

Fig. 1. Detection accuracy for the GMM-UBM method (MSR toolkit)

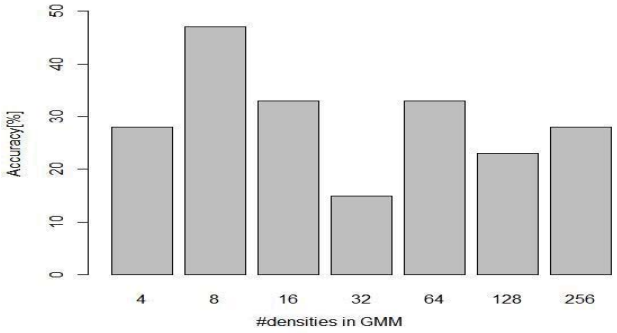

Fig. 2. Detection accuracy for the i-vectors method (MSR toolkit)

The same database, split ten times into disjoint training and evaluation sets, was used for this experiment. As in experiment #1 we varied the number of Gaussians densities per GMM (this time from 1 to 256). The results are presented in Fig. 3.

As seen in Fig. 3 the best results were obtained for the model comprising 8 Gaussian densities per GMM. The average accuracy was 61%. The reason for the difference in results obtained in experiments #1 and #3 resides in implementation details for the two toolkits used. We plan to get an insight in these implementation details in the near future.

C. *Discussion*

From the results obtained in these experiments we can say that in spectrum we can find discriminative information for each type of cry, so we can use MFCCs to characterize the cries. Also from the results we can say that the cries carry several types of information, including typical characteristics of an individual, which represents noise for the cry recognition task.

The classification accuracy obtained so far for this task are good, but not satisfactory for a system that is required to help parents understand their baby's needs. Consequently, more work is needed to answer questions such as: why are the results different when the two toolkits are used, what other methods can be used for this classification problem and, more importantly, what would be the accuracy on a larger database.

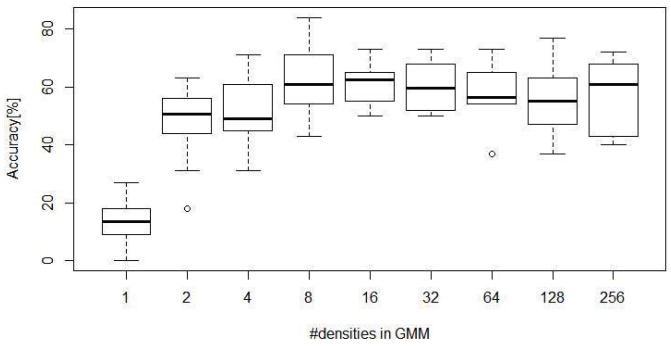

Fig. 3. Detection accuracy for the GMM-UBM method (LIUM toolkit)

V. CONCLUSIONS AND FUTURE WORK

Motivated by the fact that skilled persons such as neonatologists can distinguish between different types of baby cries, finding some patterns in each type, we developed a fully automatic system that attempts to discriminate between different types of cries.

Our results are promising and suggest that the newborn crying problem can be solved automatically, given the discriminative power of features in the spectrum for different types of cries.

ACKNOWLEDGMENT

This work was partly supported by the PN II Programme "Partnerships in priority areas" of MEN - UEFISCDI, through project no. 25/2014.

REFERENCES

[1] 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," Proc. of the 9th Conference Speech and Computer SPECOM'2004, St. Petersburg, Russia, pp. 552-557, 2004.

[2] P.S. Zeskind and B.M. Lester, "Analysis of infant crying," Chapter 8 in Biobehavioral Assessment of the Infant, L.T. Singer and P.S. Zeskind, Eds. New York: Guilford Publications Inc., 2001, pp. 149-166.

[3] P.S. Zeskind et al., "Development of Translational Methods in Spectral Analysis of Human Infant Crying and Rat Pup Ultrasonic Vocalizations for Early Neurobehavioral Assessment," Frontiers in Psychiatry, Vol. 2, Art. 56, pp. 1-16, 2011.

[4] Q. Xie, R.K. Ward, C.A. Laszlo, "Determining normal infants' level-of-distress from cry sounds," Proc. of Canadian Conf. on Electrical and Computer Engineering, pp. 1094-1096, IEEE, 1993.

[5] 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.

[6] M.S. Rusu, S.S. Diaconescu, G. Sardescu, and E. Bratila, "Database and system design for data collection of crying related to infant's needs and diseases," Proc. of the 8th Int. Conf. on Speech Technology and Human-Computer Dialogue SpeD 2015, Bucharest, Romania, 2015.

[7] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted Gaussian mixture models," in Digital Signal Processing, vol. 10, nos. 1-3. Academic Press, 2000, pp. 19-41.

[8] D. Najim, R. Dehak, P. Kenny, N. Brümmer, and P. Ouellet, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," Proc of INTERSPEECH 2009, Brighton, United Kingdom, pp. 1559-1562, 2009.

[9] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, S. Meignier, "An Open-source State-of-the-art Toolbox for Broadcast News Diarization," Proc of INTERSPEECH 2013, Lyon, France, 2013.

[10] A. Hanani, H. Basha, Y. Sharaf, S. Taylor, "Palestinian Arabic regional accent recognition," Proc. of the 8th Int. Conf. on Speech Technology and Human-Computer Dialogue SpeD 2015, Bucharest, Romania, 2015.

[11] N. Dehak, P.J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 788-798, 2011.

[12] J. T. Kwok, B. Mak, and S. Ho, "Eigenvoice Speaker Adaptation via Composite Kernel PCA," Proc of the 27th Annual Conference on Neural Information Processing Systems, Canada, 2013.

[13] S.O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1.0: a MATLAB toolbox for speaker-recognition research," in Speech and Language Processing Technical Committee Newsletter, IEEE, 2013.