Recent Experiments and Findings in Baby Cry Classification

Elena-Diana Şandru(✉), Andi Buzo, Horia Cucu, and Corneliu Burileanu

Speech and Dialogue Research Laboratory, University Politehnica of Bucharest, Bucharest, Romania
{diana.sandru, andi.buzo, horia.cucu, corneliu.burileanu}@upb.ro

Abstract. Studies have shown that newborns are crying differently depending on their need. Medical experts state that the need behind a newborn cry can be identified by listening to the cry – an easy task for specialists but extremely hard for unskilled parents, who want to act as fast as possible to comfort their baby. In this paper, we propose various experiments on a previously developed fully automatic system that attempts to discriminate between different types of cries, based on Gaussian Mixture Models. The experiments show promising results despite the difficulty of the task.

Keywords: Baby cry · Automatic newborn cry recognition · GMM-UBM K-Fold cross validation

1 Introduction

Newborns are fully dependent on adults and they are able to communicate only by crying. In this manner, the newborns express their physical and emotional state; there are several reasons for a newborn to cry, but the list can be synthesized in: hunger, tiredness, pain, discomfort, colic, eructation. As a parent it can sometimes be hard to work out which need the newborn wants you to take care of, making the process of fast need satisfaction really difficult.

Specialists (neonatologists and pediatricians) have the necessary experience to distinguish between different types of cries and identify the newborn's need. Nonetheless, medical experts consider that determining the need behind the cry is more accurate if the expert also looks at the expression of the face or the position of the body; the experience level of the listener is also essential. In addition, the newborn's cry can be used to determine whether he/she suffers from pathological diseases, thus problems can be early discovered and attended, this being 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].

The newborns' cry represents in fact a special case of human speech, being considered a short-term stationary signal (as the speech). However, the cry signal is a more random signal than a periodic one, since newborns do not have full control of the vocal tract. From functional point of view, a newborn cry is composed of four parts, in this

O. Fratu et al. (Eds.): FABULOUS 2017, LNICST 241, pp. 253–260, 2018.
https://doi.org/10.1007/978-3-319-92213-3_37

order: a sound coming from the expiration phase, a brief pause, a sound coming from the inspiration, phase followed by another pause. Obviously, there are elements that can be individualized for each child, such as sound, pause duration or even repetitions [2]. Compared to adult speech, a cry represents a complex neurophysiological act, being the outcome of an intense expulsion of air pressure in the lungs, causing oscillations of the vocal cords and leading to the formation of sound waves [3].

Abou-Abbas et al. [4] determined important audible cry component boundaries of continuous recordings collected noisy environment, in order to build a cry database, thus contributing to developing different cry based applications.

The clear advantages of identifying the need behind a newborn's cry determined several researchers to build automated systems for recognizing newborn needs using the acoustic signals of cries. However, so far most of the studies have been focused on recognizing only one need, leading to detection systems. The current paper represents recent experiments conducted within the SPLANN[1] research project, which aims to build an automatic recognition system of the newborn cries – an identification system – for five key needs: colic, discomfort, eructation, hunger and pain.

The work is based on research performed a priori, representing a continuity of the following papers: development of a fully automatic system that attempts to discriminate between different types of cries using a database created and labeled based on the Dunstan Baby Language (DBL) baby-cry classification video tutorial [5] and using the SPLANN infant crying database [6].

Both databases are labeled at expiration level, therefore the previous detection [5, 6] was in terms of expiration. The novelty of this paper emerges from a new approach – the detection for SPLANN database is performed also at cry level, namely a succession of expirations (variable), instead of a single one. In order to obtain results at cry level, a score fusion technique was used. For automatic classification, we employed a method that was successful in speaker and language recognition: the GMM-UBM (Gaussian Mixture Model - Universal Background Model). At the same time, a series of experiments have been carried out to determine to what extent the signal bandwidth and derivatives of MFCC (Mel Frequency Cepstral Coefficients) features contain discriminative information between needs, in other words, have an impact in the accuracy of recognition.

The rest of the paper is structured as follows. Section 2 describes in detail the classification methods used in the experiments. Section 3 presents the recent experiments conducted in the baby cry classification area and finally Sect. 4 is dedicated to conclusions.

2 Baby Cry Classification Methods

The classification of a baby cry based only on the audio signal represents a task similar to the classification of phonemes, closed-set speaker identification and audio-based closed-set language identification. This study addresses the problem of newborn cry

[1] SPLANN research project: http://www.softwinresearch.ro/index.php/en/research-projects/splann.

recognition by classification algorithm based on GMM, successful in speaker recognition [7]. GMMs are trained with cry recordings from different categories, labeled in advance, as seen in Fig. 1.

The training is done in two stages; in the first stage, a universal model is trained using all the recordings from all categories – the algorithm for UBM training is EM (Expectation Maximization). In the second stage, UBM model is adapted using all the recordings of a specific type of cry – the algorithm used for this adaptation is MAP (Maximum A Posteriori Probability).

The process of recognition itself is simple: each test recording is evaluated with each GMM and is scored against that particular GMM. The model for which the highest score was obtained is the winner, being used to decide which need was causing the baby to cry – namely, the cry class attached to GMM will label the given recording.

As mentioned in Sect. 1, the aim of the study was to obtain results *at cry level*, namely a succession of expirations, instead of a single one. Thus, a score fusion technique was used (as illustrated in Fig. 2). This method involves three steps: (i) obtaining a score for each need (cry class) for each expiration, (ii) summing the need scores over a sequence of N expirations and (iii) comparing the scores obtained for each need. The need with the highest score is declared to be the need for the whole sequence of expirations (for the baby cry).

An important aspect regarding this approach is deciding whether the first expirations of a cry are more discriminatory than the rest. This question arises from the fact that if the newborn's need is not satisfied quickly, he/she begins to cry hysterically and the complaints of different needs begin to resemble. To test this hypothesis we filtered the first N expirations (by varying N) of each cry and trained and evaluated the system each time with N changed.

Driven by the small dimension of Dunstan database [5], K-Fold cross validation method was implemented as validation method, such that the results to be statistically significant. K-Fold means dividing the database as follows:

- all recordings of a need are randomly sorted;
- the database per need is split into 10 equal divisions;
- during 10 experiments with the same purpose, each division will be once a test database and 9 times part of the training database;
- steps 1–3 are repeated 10 times resulting 100 datasets;
- to compute the accuracy, the averages of each iteration with 10 folds is taken. From these, the final average and standard deviation are computed and reported in the experimental test sections.

The same approach was used for SPLANN database, but since K-Fold is an exhaustive method, an arbitrary database was used to vary the way of extracting the signal features (sampling frequency, derived features) – once the most appropriate features were chosen, experiments are conducted with K-Fold to take into account their statistical significance.

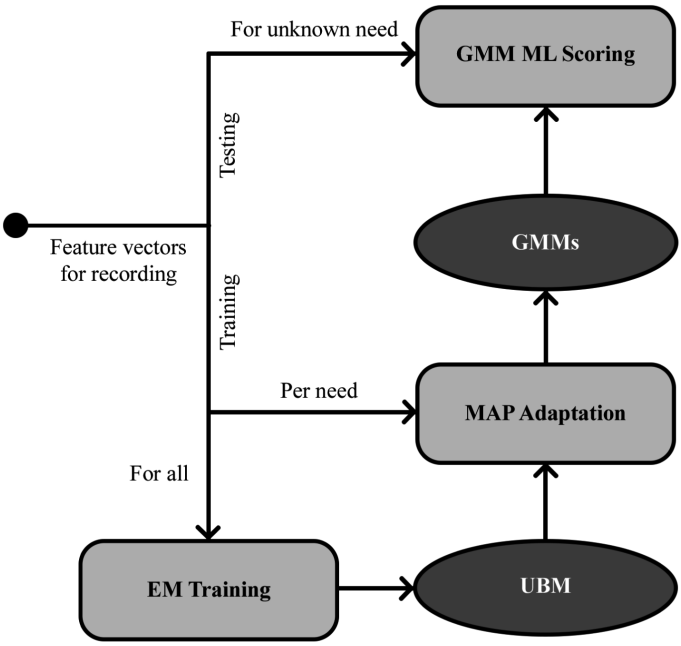

Fig. 1. GMM-based recognition scheme

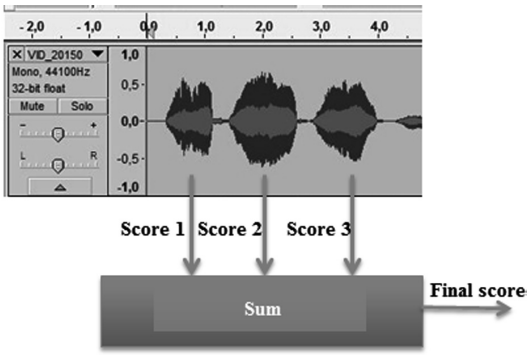

Fig. 2. Score fusion technique working principle

3 Baby Cry Classification Experiments

3.1 Baby Cry Databases and Audio Features

For the experiments presented in this paper, two databases are used. The first one – Dunstan Baby Language Database [5] has been used to prove that information about need is found in the spectrum of the signal; it contains cries for five needs: "EAIRH" – flatulence, "EH" – eructation, "HEH" – discomfort, "NEH" – hunger, "OWH" –

tiredness. The second database used was SPLANN Infant Crying Database [6], seven times bigger than the Dunstan database; the recordings of the cries were collected in the hospital with the help of a mobile phone application.

The application allowed neonatal experts to easily label the need for crying after they diagnosed it (having the possibility to look at the face expression and body position of the newborn); the data labeling validation consisted in comforting the newborn. This database contains five types of cries (eructation, discomfort, hunger, pain, colic).

To characterize the cry type we used 13 MFCC extracted from overlapped windows (a similar setup as for speech and speaker recognition); moreover, we also used their derivatives and accelerations in order to determine to what extent the signal band and MFCC derivative features contain discriminative information between needs, in other words, have a role in the accuracy of recognition. For the validation we used K-Fold permutations. The metric used to compare different scenarios was recognition accuracy: the percentage of correct classifications among all classifications performed.

3.2 Experimental Results

A. Dunstan Database – Signal Bandwidth and MFCC Features Derivatives

The experiment aimed to determine to what extent the signal bandwidth and derivatives of MFCC features contain discriminative information between needs; the experiment was conducted at expiration level, using K-Fold permutations.

Table 1 illustrates the results obtained on recognition tasks; first column highlights the sampling frequency (implicitly indicating the signal bandwidth) and numbers 13 or 39 suggest that only MFCC (the 13 coefficients) or derivatives and accelerations (a total of 39 coefficients) were used. The values presented are average values, with the confidence interval of 95% being $\pm$ 6%. Columns 2 to 6 illustrate the results obtained starting with a simple GMM-UBM (4 Gaussian densities) up to a relatively complex model (with 128 Gaussian densities).

The results presented as recognition accuracy (percentage of correctly identified needs for the tested cries) are presented in Table 1. We can observe that when the signal band is larger, the recognition accuracy is also higher – this means that there is information about need at high frequencies. Also, MFCC derivatives and accelerations contribute to the discrimination of needs.

Table 1. Recognition results on Dunstan database (in percent)

Features/#densities	4	8	32	64	128
MFCC_8kHz_13	57.5	71.0	65.9	63.3	62.4
MFCC_8kHz_39	59.1	73.5	73.4	71.1	72.8
MFCC_16kHz_13	68.9	68.8	61.5	61.4	60.4
MFCC_16kHz_39	61.1	77.5	77.4	72.6	69.5
MFCC_32kHz_13	76.8	70.1	68.3	61.5	59.4
MFCC_32kHz_39	65.3	79.5	79.1	75.5	72.3

B. SPLANN Database – Signal Bandwidth and MFCC Features Derivatives

We replicated the same experiment presented in Sect. 3.2(A) for SPLANN database, but using the arbitrary database. Table 2 shows the obtained results at expiration level and Table 3 illustrates the recognition accuracy for the entire succession of expirations for a cry, using score fusion technique. Again, when the signal bandwidth is higher, the accuracy of the recognition is also higher – this means that there is information about need at high frequencies. At expiration level, MFCC derivatives and accelerations contribute to the discrimination of needs; however, the same behavior cannot be observed at cry level. It is possible the expirations recovered due to MFCC derivatives and accelerations may not affect the decision already taken based on other expiration.

Table 2. Recognition accuracy vs features and number of Gaussian densities at expiration level

Features/#densities	4	8	32	64	128	256	512
MFCC_8kHz_13	29.7	28.4	33.4	32.4	33.8	35.5	36.2
MFCC_8kHz_39	27.7	29.1	32.4	31.7	33.4	35.4	35.8
MFCC_16kHz_13	34.0	33.1	34.9	36.1	37.8	39.2	39.1
MFCC_16kHz_39	32.1	34.7	38.0	37.7	41.7	42.6	39.9
MFCC_32kHz_13	34.3	37.8	40.4	40.0	38.7	38.2	42.2
MFCC_32kHz_39	37,8	39.5	42.7	44.1	45.3	47.2	47.3

Table 3. Recognition accuracy vs features and number of Gaussian densities at cry level

Features/#densities	4	8	32	64	128	256	512
MFCC_8kHz_13	31.3	32.8	34.3	34.3	34.3	35.8	40.3
MFCC_8kHz_39	29.9	32.8	35.8	31.3	35.8	35.8	34.3
MFCC_16kHz_13	44.8	38.8	40.3	37.3	40.3	41.8	38.8
MFCC_16kHz_39	38.8	37.3	37.3	37.3	38.8	41.8	38.8
MFCC_32kHz_13	44.8	50.7	52.2	50.7	49.3	46.3	49.3
MFCC_32kHz_39	49.3	50.7	50.7	50.7	52.2	50.7	52.2

C. SPLANN Database – First N Expirations vs Entire Cry

Motivated to find how much useful information contains the entire succession of expirations in a cry, we performed this experiment to find if the first expirations of a cry are more discriminatory than the rest. This question arises from the fact that if the newborn is not satisfied with his/her need, he/she begins to cry hysterically and the cries due to different needs begin to resemble.

We filtered the first N expirations (by varying N) of each cry and we trained and evaluated the system every time N was changed. Because we expected the difference in results to depend strongly on the choice of the test database and given the fact that using this method the database has been compressed, we used K-Fold permutations to compute a confidence interval. It should be noted that not all the cries had N expirations; in these cases, all the expirations available from the cry were taken.

Figure 3 illustrates the recognition accuracy at expiration level – it shows a slightly decreasing trend with *N*. This trend is not observed in the results at cry level presented in Fig. 4. A possible explanation may be the following: the end-of-cry expirations that have a lower score fail during the fusion to overturn the score with the first exhalations. But with these results, we can neither accept the hypothesis nor reject it. The experiment should be repeated when a larger database is available.

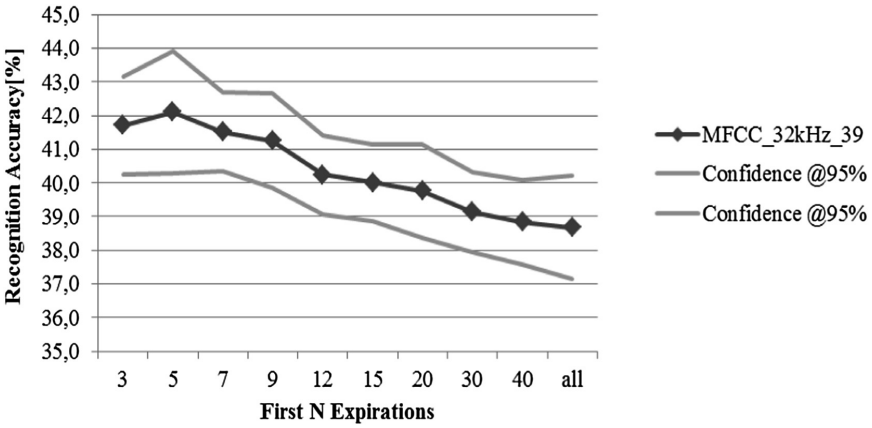

Fig. 3. Recognition accuracy at expiration level with the first N expirations

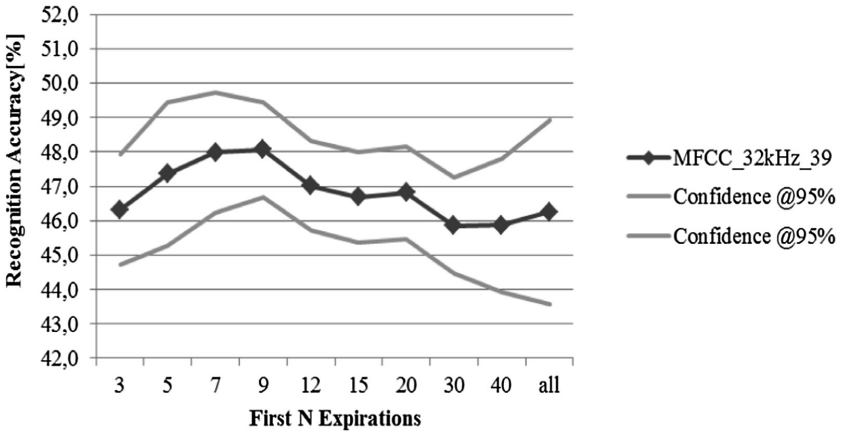

Fig. 4. Recognition accuracy at cry level with the first N expirations

4 Discussion and Conclusions

Motivated by our previous results on infant cry classification, we've extended our research on refining the previously developed automatic system that attempts to discriminate between different types of cries. The first important conclusion is that there are discriminatory elements in the signal spectrum captured in the MFCC features – confirmed by the experimental results where the accuracy much higher than the random threshold was obtained.

However, the level of discrimination is not good enough for satisfactory accuracy of the recognition system. Therefore, other features, e.g. features from time analysis, should be included in the analysis. The cry signal spectrum contains information about the need at frequencies up to 16 kHz. The features dynamics (their change over time) has a discriminatory role, as evidenced by the higher recognition accuracy when the derivatives and accelerations of the MFCC features have been used.

Future work involves an in-depth analysis of other cry features, but also of other recognition algorithms (such as Support Vector Machine). Finally, the crying recognition system must be integrated into a multimodal recognition system. This system may also contain image or video recognition elements.

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

References

1. 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: Proceedings of the 9th Conference Speech and Computer SPECOM 2004, pp. 552–557, St. Petersburg, Russia (2004)
2. Zeskind, P.S., et al.: Development of translational methods in spectral analysis of human infant crying and rat pup ultrasonic vocalizations for early neurobehavioral assessment. Front. Psychiatry **2**(Art. 56), 1–16 (2011)
3. Zeskind, P.S., Lester, B.M.: Analysis of infant crying, chapter 8 in biobehavioral assessment of the infant. In: Singer, L.T., Zeskind, P.S. (eds.), pp. 149–166. Guilford Publications Inc., New York (2001)
4. 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. J. Acoust. Soc. Am. **142**, 1318–1331 (2017)
5. Bănică, I.-A., Cucu, H., Buzo, A., Burileanu, D., Burileanu, C.: Automatic methods for infant cry classification. In: Proceedings of International Conference on Communications COMM 2016, Bucharest, Romania (2016)
6. Bănică, I.-A., Cucu, H., Buzo, A., Burileanu, D., Burileanu, C.: Baby cry recognition in real-world conditions. In: Proceedings of International Conference on Telecommunications and Signal Processing, TSP 2016, Vienna, Austria (2016)
7. Togneri, R., Pullella, D.: An overview of speaker identification: accuracy and robustness issues. IEEE Circuits Syst. Mag. **11**(2), 23–61 (2011)