Baby Cry Recognition Based on Acoustic Segment Model

Shuxian Wang, Jun Du(✉), and Yajian Wang

University of Science and Technology of China, Hefei, Anhui, China
{sxwang21,yajian}@mail.ustc.edu.cn, jundu@ustc.edu.cn

Abstract. Since babies cannot speak, they can only communicate with the outside world and express their emotions and needs through crying. Considering the variety of reasons why babies cry, it is a challenging task to accurately understand the meaning of baby crying. In this paper, we propose a baby cry recognition method based on acoustic segment model (ASM). Firstly, based on Gaussian mixtures models - hidden Markov models (GMM-HMMs), baby cry recordings are transcribed into ASM sequences composed of ASM units. In this way, different baby cry recordings are segmented in more detail, which can better capture the similarities and differences between acoustic segments. Then, by using latent semantic analysis (LSA), these ASM sequences are converted into feature vectors, and the term-document matrix is obtained. Finally, a simple classifier is adopted to distinguish different types of baby crying. The effectiveness of the proposed method is evaluated on two infant crying databases. The ASM-based approach can achieve higher accuracy compared with the approach based on residual network (ResNet). And through experiments, we analyze the reasons for the better performance of the ASM-based method.

Keywords: baby cry recognition · acoustic segment model · latent semantic analysis · deep neural network

1 Introduction

Baby cry recognition (BCR) is a task to identify the needs contained in a baby's cry [1]. Since babies do not yet have the ability to speak, crying has become the most important way for them to convey their physical and psychological needs to the outside world [2–7]. However, novice parents usually have little parenting experience. When babies cry, they are often at a loss. What's more serious is that when a baby cries because of pathological pain, if the novice parent cannot quickly and accurately understand the meaning of the baby's cry and make a wrong judgment, it is likely to miss the best time for treatment. Therefore, how to quickly understand the meaning of a baby's cry and make timely and accurate judgments is an urgent problem for every novice parent. It can be seen that the task of baby cry recognition has important research significance.

L. Zhenhua et al. (Eds.): NCMMSC 2022, CCIS 1765, pp. 16–29, 2023.
https://doi.org/10.1007/978-981-99-2401-1_2

Recently, many technical methods have been applied to the research of BCR, including traditional classifier methods such as Gaussian Mixture Models - Universal Background Models (GMM-UBM), i-vectors methods [8,9] and some methods based on deep learning: feed-forward neural networks (FNN) [2,10], time-delay neural networks (TDNN) [11], and convolutional neural networks (CNN) [12,13]. Although these methods mentioned above have achieved certain results in the recognition of baby crying, there are still some problems worthy of discussion. On the one hand, traditional methods such as GMM cannot learn deep non-linear feature transformation. On the other hand, deep learning-based methods such as CNN require a sufficient amount of data, and the difficulty of network training will increase as the number of network layers increases. In addition, it classifies infant crying by learning the feature information corresponding to the entire audio. As a result, it tends to be disturbed by longer but indistinguishable segments of the audio, while ignoring shorter but critical segments, so it cannot accurately locate the key segments that distinguish different types of infant crying.

Therefore, this paper proposes an infant cry recognition method based on the acoustic segment model (ASM), which combines the advantages of traditional methods and deep learning methods well. It can accurately mine acoustic information and segment the entire audio into more detailed segments according to whether the acoustic features have changed, so as to locate the key segments that can distinguish different categories of infant crying. ASM has been successfully applied to many tasks, such as automatic speech recognition (ASR) [14], speech emotion recognition [15,16], speaker recognition [17], music genre classification [18] and acoustic scene classification (ASC) [19]. Just as the basic building blocks of language are phonemes and grammars, baby crying signals that contain different needs of babies are also composed of fundamental units, and these fundamental units are related to each other. The proposed ASM method aims to find a universal set of acoustic units from baby cries to distinguish different types of baby cries.

The ASM framework generally consists of two steps, namely initial segmentation and iterative modeling. In the initial segmentation step, there are many different segmentation methods to obtain the basic acoustic units, such as maximum likelihood segmentation [14,20], even segmentation [15], K-means clustering algorithm [21], etc. The segmentation method used in this paper is GMM-HMMs, that is, each type of baby crying is modeled by GMM-HMMs [22–24]. Specifically, according to the similarities and differences of acoustic characteristics, the segments with similar acoustic characteristics are grouped together and marked with the same hidden state. Each hidden state corresponds to an ASM unit. In this way, through the initial segmentation, each baby cry recording is divided into variable length segments, so that we get the initial ASM sequences. Then, for iterative modeling, each ASM unit is modeled by a GMM-HMM and then baby cries are decoded into a new sequence of ASM units. After transcribing a baby cry into an ASM sequence, each baby cry is composed of ASM units, which is similar to a text document composed of terms. Therefore, we can use

latent semantic analysis (LSA) to generate the term-document matrix. Each column of the matrix is a feature vector of a baby cry recording, and then these feature vectors are sent to the backend classifier.

The remainder of the paper is organized as follows. In Sect. 2, we introduce the proposed model and method used in baby cry recognition. In Sect. 3, experimental results and analysis are presented. Finally, we make the conclusions of this study and summarize our work in Sect. 4.

2 Method

For BCR, this paper proposes an ASM-based analysis method. The framework of the method is shown in Fig. 1. For the training data, through the two steps of initial segmentation and iterative modeling, they are all transcribed into ASM sequences. At the same time, the acoustic segment model generated in the iteration can be used to transcribe the test data into ASM sequences. In this way, each acoustic recording is transcribed into a sequence composed of ASM units, which is similar to a text document composed of terms. Therefore, text classification methods widely used in the field of information retrieval, such as LSA, can be used to analyze this problem. Through LSA and singular value decomposition (SVD) [25], an ASM sequence can be converted into a vector, so that the ASM sequences transcribed from all training data can be mapped to a term-document matrix. Each sample in the test set is processed in the same way. After the above processing, we can get the feature vector corresponding to each sample in the training and test sets, and then send these vectors to the backend DNN for classification.

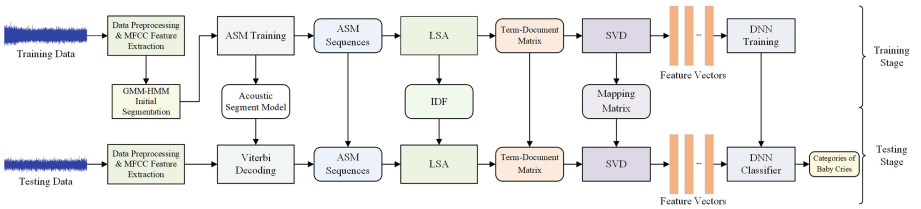

Fig. 1. ASM-based system framework.

2.1 Acoustic Segment Model

The main function of the acoustic segment model is to convert a baby cry recording into a sequence composed of basic acoustic units, just like a sentence is composed of words. The ASM method usually consists of two stages, namely initial segmentation and model training.

Initial Segmentation. The initial segmentation affects the result of the ASM method, so it is a very critical step. Many methods have been proposed to do the segmentation. Considering that GMM-HMMs achieve outstanding performance in ASR and can well explore the boundaries of acoustic feature changes, we use GMM-HMMs to perform initial segmentation.

First of all, each baby cry is modeled by GMM-HMMs, and the HMM is a left-to-right topology. However, considering that similar baby cries may occur at different time periods, we add a swivel structure to the topology, so that similar frames can be represented by the same hidden state. Assuming that there are B kinds of baby crying in the database, each kind of baby crying is modeled by a GMM-HMM with M hidden states. And then, through the Baum-Welch algorithm [26], we can update the parameters of GMM-HMMs. Through decoding, each baby cry recording is transcribed into a sequence composed of hidden states, and each hidden state is corresponding to a segment of the baby cry recording. Thus, the $C = B \times M$ hidden states are used as the corpus to initialize each baby cry recording as an ASM sequence.

Model Training. After completing the initial segmentation, each baby cry recording is converted into a sequence composed of ASM units. In the model training step, first, we use the GMM-HMM with a left-to-right HMM topology to model each ASM unit. Then, the Baum-Welch algorithm is adopted to update the parameters of the GMM-HMM model. Next, we use the Viterbi decoding algorithm to transcribe the training set data into new ASM sequences. These new ASM sequences are used as new labels for the training recordings in the next iteration of model training. The above process is repeated until the ASM sequences corresponding to the training data converge stably.

2.2 Latent Semantic Analysis

After the processing of the above steps, each baby cry recording is transcribed into a sequence composed of ASM units, which is like a text document composed of terms. Therefore, in view of the outstanding performance of LSA in the field of text processing, we use LSA for analysis, that is, through the LSA method, the correspondence between ASM units and baby cry recordings can be described by a term-text document matrix. Each column of the matrix corresponds to each baby cry recording that has been transcribed, and each row corresponds to one ASM unit or two adjacent ASM units in the ASM sequence. Therefore, if there are C elements in the baby cry corpus, then the dimension of the vector in each column of the matrix is $D = C \times (C + 1)$.

Similar to the processing method in the field of information retrieval, the value of each element in the matrix is determined by term frequency (TF) and inverse document frequency (IDF) [27]. TF reflects the frequency of a word in the current text, and IDF reflects the frequency of the word in all texts. If a word has a high probability of appearing in all texts, even if it appears many times in a certain text, its importance to the text is also very low. Therefore,

only by integrating TF and IDF, can it more accurately reflect the importance of a word to a text. The formula for calculating the TF of the i-th ASM term in the j-th baby cry recording is as follows:

$$TF_{i,j} = \frac{x_{i,j}}{\sum_{d=1}^{D} x_{d,j}}, \tag{1}$$

where $x_{i,j}$ is the number of times the i-th term appears in the ASM sequence corresponding to the j-th baby cry recording. The IDF calculation formula is as follows:

$$IDF_i = \log \frac{Q+1}{Q(i)+1}, \tag{2}$$

where Q is the number of baby cry recordings in the training set, and $Q(i)$ is the number of texts in which the i-th term has appeared. In this way, each element of the matrix W is defined as follows:

$$w_{i,j} = TF_{i,j} \times IDF_i. \tag{3}$$

Due to the use of bigrams with sparsity problems, the term-document matrix W with dimension $D \times Q$ is sparse. We can use the SVD to reduce the dimension of the matrix W, as follows:

$$W = U \Sigma V^T. \tag{4}$$

The matrix W is decomposed into the product of three matrices: the left-singular $D \times D$ matrix U, the diagonal $D \times Q$ matrix Σ and the right-singular $Q \times Q$ matrix V. Among them, the diagonal elements of matrix Σ are the singular values of matrix W, and these singular values are arranged from large to small. We take the first t singular values and the first t rows of the matrix U to form a mapping space U_t, and then multiply this matrix with the original matrix W to get a new matrix W_t after dimensionality reduction. And W_t is used as the training data of the backend classifier. The value of t is determined based on the percentage of the sum of the squares of the singular values.

For the test data, the processing method is similar to the above steps of the training data, but what we need to pay attention to is that the test data needs to use the IDF and matrix U_t obtained in the training phase to obtain the matrix W_t^{test}.

2.3 DNN Classifier

The classifier used in this paper is DNN, which is used to distinguish different types of baby crying. Since the feature vectors of baby crying extracted by the ASM method are easy to distinguish, a simple DNN structure is adopted.

Table 1. The details of two baby crying databases.

Database	A	B
Recording Scene	Home	Hospital
Number of Categories	6	2
Training Set	8.03 h	8.54 h, 2020.02-2020.12
Validation Set	3.44 h, seen babies	N/A
Test Set	2.41 h, unseen babies	1.34 h, 2021.01

3 Experiments and Analysis

3.1 Database and Data Preprocessing

There are few high-quality infant crying databases that have been published, which brings certain challenges to the research of baby cry recognition. In this study, we adopt two infant crying databases, which were recorded at home and in the hospital, respectively, to evaluate our method. It is worth mentioning that the baby crying data recorded at home was annotated by parents based on their own experience or subsequent processing response, while the data recorded in the hospital was annotated by hospital pediatric experts. In addition, in the data set recorded in the hospital, some of the cries were recorded when the baby was given injections due to illness, so these cries were labeled "pain". Therefore, for the crying data recorded in the hospital, we conduct a two-class analysis of "pain" and "non-pain". The baby crying data recorded at home has six types of labels: "hungry", "sleepy", "uncomfortable", "pain", "hug", and "diaper", which are analyzed in six categories.

Before analyzing these baby crying data, we need to preprocess the data. First, down-sampling processing is performed to unify the sampling rate of all baby cry recordings 16000 Hz. Then, we apply voice activity detection (VAD) to detect the start and end of the silent interval in a baby cry recording according to the difference between the energy of the silent interval and the non-silent interval. The VAD method is used to remove silent redundancy and expand the number of effective samples. Finally, all the non-silent intervals of a baby cry recording are combined and then divided into 10-second segments.

After the above data preprocessing, the detailed information of the two infant crying databases is shown in Table 1. For database A, it was recorded at home and consists of crying recordings of 65 babies aged 0–6 months, and a certain amount of crying data is collected for each baby. We randomly select the crying data of 10 babies as the test set, and the rest of the crying data are mixed together and divided into training set and validation set. Therefore, the babies in the validation set have been seen in the training set, and the babies in the test set have not been seen in the training set. For database B, it was recorded in the hospital. The recording time is from February 2020 to January 2021. We select

the crying data recorded in January 2021 as the test set, and the rest of the crying data as the training set. Since there are many babies in the delivery room of the hospital, and considering the protection of the privacy of the newborn, this part of the data cannot record the ID of the baby. Therefore, the database B is not divided according to the baby ID but the recording month, and because the neonatal hospitalization time is limited and the infants hospitalized in the ward are highly mobile, only one or two cry data are recorded for each baby, so it can be considered that the crying in the test set comes from the infants that have not been seen in the training set.

3.2 Ablation Experiments

Baseline System. Residual network (ResNet) [28] has been successfully used in the research of infant cry recognition [29], and has performed well on audio classification tasks in recent years, so ResNet is adopted as the baseline system.

In our previous work, ResNet excelled in the acoustic scene classification (ASC) task [30]. Considering that these two tasks are similar, we adopt the ResNet structure used in the ASC task [30], and then the ResNet is trained on these two baby cry databases. The log-Mel spectrogram is used as the input to the baseline system. To generate log-Mel spectrogram, a short-time Fourier transform (STFT) with 2048 FFT points is applied, using a window size of 2048 samples and a frameshift of 1024 samples, and log-Mel deltas and delta-deltas without padding are also computed.

ASM-DNN Baby Cry Recognition Model. The structure and principle of the ASM-DNN baby cry recognition model have been introduced in Sect. 2, which uses MFCC as the input feature. 60-dimensional Mel-frequency cepstral coefficients (MFCC) features are extracted by using a 40-ms window length with a 20-ms window shift to train GMM-HMMs. The DNN we used has 3 hidden layers, each with 512 neurons, and a fixed dropout rate of 0.1. The parameters of the DNN are learned using the SGD [31] algorithm, and the initial learning rate is set to 0.1. We perform ablation experiments on database A to explore the impact of two important parameters in the ASM-DNN model (the number of ASM units and the percentage of dimensionality reduction after singular value decomposition (SVD)) on the recognition accuracy.

- **Number of ASM units**

Obviously, if the number of ASM units is too small, the differences between different baby cry clips cannot be captured well, but if the number of ASM units is too large, it may cause overfitting problems, so we need to find a suitable value through experiments.

The number of ASM units is equal to the total number of hidden states, that is, assuming that there are B kinds of baby crying in the database, and each kind of baby crying is modeled by a GMM-HMM with M hidden states, then the number of ASM units is $C = B \times M$. As mentioned earlier, database

A contains six types of crying. We adjust the number of ASM units by adjusting M. Table 2 shows the experimental results with different number of ASM units. It is worth noting that in these experiments, the first 70% of the singular values are retained after singular value decomposition. It can be observed from Table 2 that the best result achieved when each kind of baby crying is modeled with 6 hidden states. Meanwhile, an accuracy of 29.99% is achieved on the test set.

Table 2. Performance comparisons with different ASM units.

Hidden States	ASM Units	The Accuracy of Validation Set
4	24	45.45%
5	30	46.33%
6	36	**47.21%**
7	42	46.57%

- **Dimensionality Reduction in SVD**

As mentioned earlier, the term-document matrix obtained by LSA is sparse, so we can reduce the dimensionality of the matrix by retaining the largest singular values after SVD. The dimension of the new matrix obtained after dimensionality reduction is determined by the percentage of the sum of squares of singular values. Keeping 36 ASM units, we adjust the percentage of dimensionality reduction and the results are presented in Table 3. We can observe from Table 3 that when the percentage is set to 70%, the model achieves the highest accuracy on the validation set. At this time, the model's recognition accuracy of the test set is 29.99%.

Table 3. Performance comparisons with different reduced dimensions in SVD.

Percentage	The Accuracy of Validation Set
60%	47.05%
70%	**47.21%**
80%	46.81%

In summary, the optimal number of hidden states for modeling each type of baby crying is 6, so for databases A and B, the optimal number of ASM units is 36 and 12, respectively. Meanwhile, the best SVD dimension reduction dimension is 70%.

3.3 Overall Comparison

Overall Comparison on Baby Crying Database A. The results of these two approaches on database A are shown in Table 4. By comparing ResNet and ASM-DNN, we can observe that better results are obtained for ASM-based approach. Specifically, the recognition accuracy of crying for both seen babies and unseen babies is improved by adopting the ASM-based approach. In addition, it can be seen that the recognition accuracy of these two approaches on the test set is 28.49% and 29.99%, respectively, and neither exceeds 30%. Considering that the individual differences of babies will cause different babies' crying to be different, this may cause the model to have lower accuracy when recognizing the crying of babies that have not been seen in the training set.

Table 4. The accuracy comparisons between ResNet and ASM-DNN on database A.

System	Validation Set	Test Set
ResNet	39.58%	28.49%
ASM-DNN	**47.21%**	**29.99%**

Overall Comparison on Baby Crying Database B. The results of these two approaches on database B are shown in Table 5. Compared with the ResNet-based approach, although the proposed ASM-based approach has a slightly lower recognition accuracy of pain crying, however, for non-pain crying, the performance of the ASM-based approach is significantly improved. Specifically, the recognition accuracy of non-pain crying is improved from 60.73% to 69.09% by adopting the ASM-based approach, that is, the performance can be improved by 8.36%. Therefore, for the entire test set, compared with ResNet, the ASM-based approach improves the recognition accuracy from 66.46% to 71.01%, which is an increase of about 5% points. Obviously, similar to the experimental results of database A, the ASM-based approach also performs better on database B, which further demonstrates the effectiveness of the ASM-based approach.

Table 5. The accuracy comparisons between ResNet and ASM-DNN on database B.

System	Pain Crying	Non-pain Crying	Overall Test Set
ResNet	**74.04%**	60.73%	66.46%
ASM-DNN	73.56%	**69.09%**	**71.01%**

3.4 Results Analysis

Spectrogram Analysis. Figures 2 and 3 show two examples of pain crying and non-pain crying from data set recorded in the hospital. The two recordings

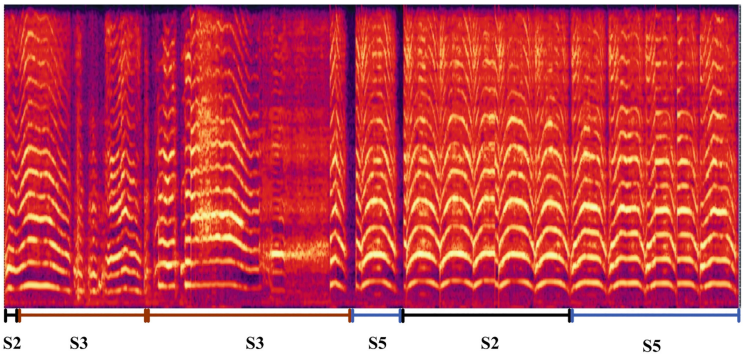

Fig. 2. The spectrogram and ASM sequence of an example recording of pain baby crying. This example was misclassified by ResNet as the non-pain baby crying but correctly classified by our ASM-DNN approach.

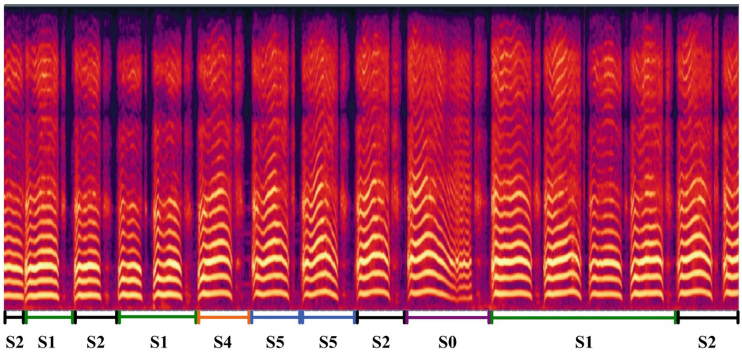

Fig. 3. The spectrogram and ASM sequence of an example recording of non-pain baby crying. This example was misclassified by ResNet as the pain baby crying but correctly classified by our ASM-DNN approach.

are confused by the ResNet, but the ASM-based model can correctly distinguish these two kinds of crying. For a more intuitive analysis, we show the results of transcribing these two cry recordings into ASM sequences through the ASM method. The ASM units are named from S0 to S11. It can be observed that similar parts in the spectrograms are represented by the same ASM units, such as S2. At the same time, the different parts can be captured by the ASM units such as S3 for pain crying and S0 for non-pain crying. It can be seen that by using the ASM-based method, we can capture the differences between acoustic segments in more detail, and then distinguish different types of baby crying more accurately. Therefore, the result of the ASM-based method is better than that of the ResNet.

Comparison of Model Judgment Results and Expert Audiometry Results. In the database of baby crying recorded by the hospital (that is, database B), we randomly select 300 pieces of data, and invite three experienced pediatric experts to conduct audiometry, and each expert conduct audiometry on 100 cries. The results of the experts' audiometry can be regarded as the "ceiling" of the recognition accuracy of the model. The audiometric results are shown in Table 6. It can be seen from Table 6 that the accuracy rates of the three experts' audiometry are 66%, 72%, and 62%, respectively. At the same time, it can be found from Table 5 that the recognition accuracy of the ASM-DNN model is 71.01%. Obviously, our model is very close to the best recognition accuracy among the three pediatric experts, which further proves the effectiveness of the ASM-DNN method.

Table 6. The results of the experts' audiometry.

Audio No.	The Accuracy of the Experts' Audiometry
1–100	66%
101–200	72%
201–300	62%

Comparative Analysis of Experimental Results on Two Databases. By observing the previous experimental results, it can be found that the recognition accuracy of the crying data recorded in the hospital is higher than that of the crying data recorded at home. The main reasons are as follows:

First of all, the crying data recorded at home is analyzed in six categories, while the crying data recorded in the hospital is analyzed in two categories. It should be noted that the more categories, the greater the uncertainty. Hence the recognition accuracy of baby crying will be correspondingly improved when only two classes are needed to be predicted.

Secondly, the category labels of baby crying data collected in the hospital are marked by experienced pediatric experts. The experts mark the crying based on their years of experience, combined with the baby's facial expressions, movements, breathing state, and the intensity of crying. However, the category labels of baby crying data recorded at home are marked by parents. Pediatric experts have more experience, so the category labels of the crying data marked by them are more accurate and reliable, which makes the recognition accuracy for data collected in the hospital is also higher.

4 Conclusions

In this study, we propose an ASM-based analysis method for baby cry recognition. We first transcribe all baby cry recordings into ASM sequences composed of ASM units through the two steps of initial segmentation and iterative modeling,

so that the similarities and differences between the segments of baby cry recordings can be well captured. Then, using LSA and SVD, a dimensionality-reduced term-document matrix is obtained. Finally, a classifier with a relatively simple structure can be used in the backend to achieve the purpose of identifying baby crying. Experiments conducted on two databases show that the ASM combined with a simple DNN classifier achieves better results than ResNet for baby cry recognition, which demonstrates the effectiveness of the ASM-based model.

References

1. Drummond, J.E., McBride, M.L., Wiebe, C.F.: The development of mothers' understanding of infant crying. Clin. Nurs. Res. **2**(4), 396–410 (1993)
2. Garcia, J.O., Garcia, C.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, pp. 3140–3145 (2003)
3. Rusu, M.S., Diaconescu, Ş.S., Sardescu, G., Brătilă, E.: Database and system design for data collection of crying related to infant's needs and diseases. In: 2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 1–6 (2015)
4. Wasz-Höckert, O., Partanen, T.J., Vuorenkoski, V., Michelsson, K., Valanne, E.: The identification of some specific meanings in infant vocalization. Experientia **20**(3), 154–154 (1964)
5. Orlandi, S., et al.: Study of cry patterns in infants at high risk for autism. In: Seventh International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (2011)
6. Farsaie Alaie, H., Tadj, C.: Cry-based classification of healthy and sick infants using adapted boosting mixture learning method for gaussian mixture models. Model. Simul. Eng. **2012**(9), 55 (2012)
7. Chittora, A., Patil, H.A.: Classification of pathological infant cries using modulation spectrogram features. In: The 9th International Symposium on Chinese Spoken Language Processing, pp. 541–545 (2014)
8. Bănicğ, I.A., Cucu, H., Buzo, A., Burileanu, D., Burileanu, C.: Baby cry recognition in real-world conditions. In: 2016 39th International Conference on Telecommunications and Signal Processing (TSP), pp. 315–318 (2016)
9. Bănică, I.A., Cucu, H., Buzo, A., Burileanu, D., Burileanu, C.: Automatic methods for infant cry classification. In: 2016 International Conference on Communications (COMM), pp. 51–54 (2016)
10. Abdulaziz, Y., Ahmad, S.M.S.: Infant cry recognition system: a comparison of system performance based on mel frequency and linear prediction cepstral coefficients. In: 2010 International Conference on Information Retrieval & Knowledge Management (CAMP), pp. 260–263 (2010)
11. 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: 9th Conference Speech and Computer, pp. 552–557 (2004)

12. Chang, C.Y., Li, J.J.: Application of deep learning for recognizing infant cries. In: 2016 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), pp. 1–2 (2016)
13. Yong, B.F., Ting, H.N., Ng, K.H.: Baby cry recognition using deep neural networks. In: World Congress on Medical Physics and Biomedical Engineering 2018, pp. 809–813 (2019)
14. Lee, C.H., Soong, F.K., Juang, B.H.: A segment model based approach to speech recognition. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 501–502 (1988)
15. Lee, H.Y., et al.: Ensemble of machine learning and acoustic segment model techniques for speech emotion and autism spectrum disorders recognition. In: INTERSPEECH, pp. 215–219 (2013)
16. Zheng, S., Du, J., Zhou, H., Bai, X., Lee, C.H., Li, S.: Speech emotion recognition based on acoustic segment model. In: 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), pp. 1–5 (2021)
17. Tsao, Y., Sun, H., Li, H., Lee, C.H.: An acoustic segment model approach to incorporating temporal information into speaker modeling for text-independent speaker recognition. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4422–4425 (2010)
18. Riley, M., Heinen, E., Ghosh, J.: A text retrieval approach to content-based audio retrieval. In: International Society for Music Information Retrieval (ISMIR), pp. 295–300 (2008)
19. Bai, X., Du, J., Wang, Z.R., Lee, C.H.: A hybrid approach to acoustic scene classification based on universal acoustic models. In: Interspeech, pp. 3619–3623 (2019)
20. Svendsen, T., Soong, F.: On the automatic segmentation of speech signals. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 77–80 (1987)
21. Hu, H., Siniscalchi, S.M., Wang, Y., Bai, X., Du, J., Lee, C.H.: An acoustic segment model based segment unit selection approach to acoustic scene classification with partial utterances. In: INTERSPEECH, pp. 1201–1205 (2020)
22. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **77**(2), 257–286 (1989)
23. Su, D., Wu, X., Xu, L.: GMM-HMM acoustic model training by a two level procedure with Gaussian components determined by automatic model selection. In: 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4890–4893 (2010)
24. Karpagavalli, S., Chandra, E.: Phoneme and word based model for tamil speech recognition using GMM-HMM. In: 2015 International Conference on Advanced Computing and Communication Systems, pp. 1–5 (2015)
25. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular value decomposition and principal component analysis. In: A Practical Approach to Microarray Data Analysis, pp. 91–109 (2003)
26. Elworthy, D.: Does Baum-Welch re-estimation help taggers?. arXiv preprint cmp-lg/9410012 (1994)
27. Hull, D.: Improving text retrieval for the routing problem using latent semantic indexing. In: SIGIR1994, pp. 282–291 (1994)
28. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
29. Xie, X., Zhang, L., Wang, J.: Application of residual network to infant crying recognition. J. Electron. Inf. Technol. **41**(1), 233–239 (2019)

30. Hu, H., Yang, C.H.H., Xia, X., et al.: A two-stage approach to device-robust acoustic scene classification. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 845–849 (2021)
31. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of COMPSTAT 2010, pp. 177–186 (2010)