Analysis of Time-Averaged Feature Extraction Techniques on Infant Cry Classification

Aditya Pusuluri(✉), Aastha Kachhi, and Hemant A. Patil

Speech Research Lab, DA-IICT Gandhinagar Gujarat, Gandhinagar, India
{aditya_pss,aastha_k,hemant_patil}@daiict.ac.in

Abstract. Classification of infant cry into normal and pathological cries is a socially relevant research problem for a long time. Crying is the *only* means that an infant use for communication. The state-of-the-art feature vectors, such as Short-Time Fourier Transform (STFT) representations and Mel Frequency Cepstral Coefficients (MFCC) have been earlier reported for this task. However, *quasi-periodic* sampling of vocal tract spectrum by high pitch-source harmonics of infant cry results in poor spectral resolution in STFT based spectrum and hence, these feature vectors could not produce satisfactory performance. In this work, we compare the performance of various time-averaged feature extraction techniques of window sizes 20, and 55 ms with three different classifiers, namely, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Random Forest (RF). The experiments in this work are performed using the *10*-fold stratified cross-validation on standard and statistically meaningful *Baby Chillanto* dataset using various state-of-the-art features vectors. It was observed that the time-averaged dynamic MFCC feature vector gives a classification accuracy of 98.48%. Furthermore, the performance of the proposed feature vectors was also studied using the confusion matrix and found to be better than other features, such as LFCC and CC.

Keywords: Infant cry classification · MFCC · LFCC · Cepstral coefficients · Time average · KNN · Random forest · SVM

1 Introduction

Crying is the only mode of communication for an infant to convey information to the parents or caregivers. The cry of an infant can be meant for many reasons, which indicate the emotional, physical, and pathological needs of infants. The exact reasoning behind the infant's cry is difficult to understand for inexperienced mothers and caregivers. Hence, the infant cry classification system can be used for the early detection and diagnosis of the infant's condition. Research has found that there is a typical pattern associated with various kinds of crying and hence, the infant cry classification problem can also be seen as a pattern classification problem. Fingerprint-based biometrics [9] were developed apart from

S. R. M. Prasanna et al. (Eds.): SPECOM 2022, LNAI 13721, pp. 590–603, 2022.
https://doi.org/10.1007/978-3-031-20980-2_50

cry-based identification [17] to prevent the infant mortality rate due to vaccine-preventable diseases and malnutrition.

The initial work on the infant cry started in the 1940s [10,13]. Later, in the 1960s, four types of infant cries were identified [22]. Ten distinct cry modes were identified based on the variation of fundamental frequency (F_0) and its harmonics from the narrowband spectrogram by Xie et al. [24]. This study was extended from normal infant cry to pathological infant cry, where *dysphonation* and *hyperphonation* cry modes were found to be correlated with the pathological cry [18]. Despite the interest in the prospects offered by the study of infant cry in the early diagnosis, scientific work was not restricted to this area alone. There has been research on categorizing the cry since the 1960s [23]. While the previous works were based on a manual study by experts and doctors, recent advances in automation and machine learning have opened the doors for automating the detection of any discomfort in the infant's cry. State-of-the-art cepstral features, such as Mel Frequency Cepstral Coefficients (MFCC) are also recently used for cry classification tasks using a Gaussian Mixture Model (GMM) classifier [1,14], using fuzzy logic based classifier [20], decision tree, Support Vector Machine, boosted tree [16], feedforward network [11]. Another state-of-the-art feature vector, namely, Linear Frequency Cepstral Coefficients (LFCC) is also used for the classification task with the k-Nearest Neighbors classifier [7,8]. However, there hasn't been a lot of work done on a comparative study of normal *versus* pathological cry classification among MFCC, LFCC, and Cepstral Coefficients feature extraction techniques.

In this work, we present a comparative study among multiple feature extraction techniques, such as MFCC, LFCC, and CC on different window sizes combined with various classifiers, namely, k-Nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machine (SVM). MFCC being the state-of-the-art feature vector replicates the hearing mechanism of the human ear, i.e., inducing non-linear characteristics in tone perception. LFCC is a feature extraction technique similar to the MFCC, where the Mel filterbank is replaced by a linear filterbank. The LFCC is found to capture information at higher frequencies better than the MFCC [8]. It is observed that the performance of the classifier with the MFCC feature vector is better than the LFCC and CC [8]. We have used various classification algorithms, namely, SVM, RF, and KNN. Instead of using the MFCC, LFCC, and CC features as it is, we averaged the feature matrix across the time-axis as most of the cepstral information of the sound wave is captured in the first 13–14 indexes of coefficient values.

The rest of this paper is organized as follows. Section 2 presents the proposed work on time averaging feature extraction. Section 3 describes the standard and statistically meaningful Baby Chillanto database. The experimental results and the analysis of the results are presented in Sect. 4. Finally, Sect. 5 concludes the paper along with potential future research work.

2 Proposed Work

In this work, we analyze multiple combinations of feature extraction techniques and classifiers. We also compare different feature extraction techniques and the

effect of averaging of feature extraction matrix on the classification accuracy. The work is done considering 2 window sizes for the feature extraction technique: 20 and 55 ms. The 20 ms is the default window size for the STFT function and 55 ms is the default window size for the MFCC function using the Librosa toolkit. Apart from the above explanation, 20 and 55 ms are selected as this reflects the clear differences in the effect of increasing the window size on the classifiers.

2.1 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is one of the state-of-the-art feature extraction techniques. The speech signal is a time-varying signal and hence, when analyzed for a short-time period, it acts as a stationary signal. One way of short-time signal analysis is by employing MFCC, which aims to develop segmental features from audio signals. The procedure for obtaining MFCC is shown in Fig. 1. The feature vector obtained after the MFCC feature extraction technique for an audio file is a 2-D array or a matrix. To perform a short-term analysis, we frame block the audio signal into different segments called *frames* with each segment having a 20 ms length with an overlap of 10 ms in general. In order to avoid the introduction of noise at higher frequency stages, we use windowing after framing to eliminate the abrupt chopping of the signal. In general, Hamming or Hanning windows are used as they result in reasonable side lobes widths with the desired main lobe width [2]. Next, Fast Fourier Transform (FFT) is used to convert the signal in the time-domain to the frequency-domain. The results are then passed through the Mel filterbank to change the frequency into the Mel scale. The conversion of frequency into the Mel scale is done by [12]:

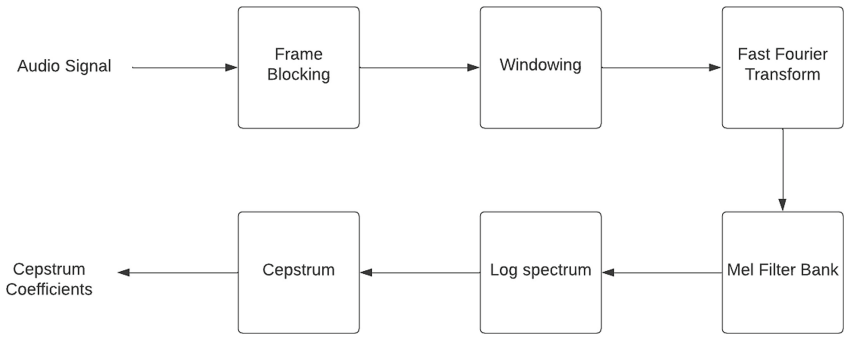

Fig. 1. Block Diagram of MFCC feature extraction [6].

$$Mel(f) = 2595 * log(1 + f/700).$$

The Mel scale filterbank is a bandpass triangular filterbank followed by a logarithmic function. This decreases the resolution at the higher frequencies and improves the resolution at lower frequencies. Later, the spectrum is passed through the DCT block to convert the spectrum to cepstrum and to decorrelate

the sub-band energies from the frequency filterbank. After obtaining the coefficients, these coefficients are termed *static* MFCC feature vectors. Later, upon applying the first-order and second-order difference (i.e., numerical approximation to the derivative operator) on the static feature vector, we get a dynamic MFCC feature vector. The dynamic features track the rate of change in features (in particular, cepstral trajectory) w.r.t time. These dynamic features can be useful, however, also redundant sometimes.

In the static MFCC feature extraction technique, 13 coefficients are considered, a window length of 20 and 55 ms, a hop length of 10 and 15 ms respectively, a minimum frequency 100 Hz, and a maximum frequency of sampling rate/2. In dynamic MFCC, 39 coefficients are considered, a window length of 20 and 55 ms, a hop length of 10 and 15 ms, a minimum frequency 100 Hz, and a maximum frequency of sampling rate/2.

2.2 Linear Frequency Cepstral Coefficients (LFCC)

LFCCs are another state-of-the-art feature extraction technique widely used. The procedure is similar to that of MFCC, here the Mel filterbank is replaced with a linear filterbank. Due to the presence of a linear frequency filterbank, the resolution is better at higher frequencies compared to the MFCC as here the spacing is not logarithmic but is linear. Hence, it captures details better at higher frequencies compared to MFCC. Further, for both MFCC and LFCC, DCT does the job of feature decorrelation, energy compaction, and dimensionality reduction of the feature vector.

Here, 13 coefficients are considered, a window length of 20 and 55 ms, a hop length of 10 and 15 ms, a minimum frequency 100 Hz, and a maximum frequency of sampling rate/2.

2.3 Cepstral Coefficients (CC)

In this technique, there is no application of any filterbank and extract the features skipping the filterbank procedure present in the MFCC and LFCC feature sets.

13 coefficients are considered, a window length of 20 and 55 ms, a hop length of 10 and 15 ms, a minimum frequency 100 Hz, and a maximum frequency of sampling rate/2.

2.4 Time Averaging of Features

The sound files of normal *versus* pathology infant cry classification are recordings of cries and they don't contain any time-specific information and our the is to classify the infant cry but not detect the infant cry. Hence, the temporal axis in the extracted features of MFCC, LFCC, and CC doesn't contain much information, and averaging them doesn't lead to information loss, which can be proved from the classification results obtained. This time-averaging technique

helps to overcome the computational complexity while obtaining a good classification accuracy. Another explanation to justify the averaging of feature vectors across time is that as the window size is increased in the feature extraction technique, the time resolution decreases and the frequency resolution increases, and the average classification accuracy increased for every feature extraction technique. This implies that for the infant cry classification, the information across the frequency axis of the matrix obtained from the feature extraction technique is more informative than the information obtained from the time-axis.

3 Experimental Setup

3.1 Dataset Used

Baby Chillanto dataset is used in this work. It was developed by the recordings conducted by medical doctors, which is a property of NIAOE-CONACYT, Mexico [21]. Each cry signal was segmented into one-second duration (which represents one sample), and is grouped into five categories. Two groups were formed for binary classification of healthy *versus* pathology. Healthy cry signals include three categories, namely, normal, hungry, and pain resulting in 1049 cry samples. Pathology cry signals include two categories, namely, asphyxia and deaf resulting in 1219 cry samples. Table 1 shows the statistics of Baby Chilanto database. The normal class consists of 1038 samples and the pathology class consists of 1229 samples. 70% of the data is used for training, and 30% of the data is used for testing.

Table 1. Statistics of the Baby Chillanto dataset used [21].

Class	Category	# Utterances
Healthy	Normal	507
	Hungry	350
	Pain	192
Pathology	Asphyxia	340
	Deaf	879

3.2 Classifier Parameters

Support Vector Machines (SVM): It is a non-probabilistic binary linear classifier as it assigns any new data sample directly to one of the classes. The SVM is based on discriminative training and it gives an optimal hyperplane in the higher-dimensional feature space than the dimension of the original feature vector, given labeled training samples that categorize new examples [3]. In particular, SVM is based on Cover's theorem on the separability of patterns, i.e., the patterns that are nonlinearly separable in low-dimensional feature space become

linearly separable in the high-dimensional feature space by using a suitable kernel function [5]. Here the classification is done using a decision boundary. Table 2 specifies the best parameters obtained for SVM with a linear kernel using grid search algorithm.

Table 2. Parameter tuning for SVM using grid search method for a window size of 55 ms.

Parameter	Static MFCC	Dynamic MFCC	LFCC	CC
C	0.1	1	10	100

K Nearest Neighbours (KNN): KNN is a well-known pattern recognition method that helps to classify binary or multiple classes which are having its own label vectors. KNN classifier determines the class based on the concept of majority voting of the nearest neighbors. The nearest neighbors are measured using a distance metric. The Euclidean distance metric is one of the most commonly used distances [3]. Here the classification is done using the concept of clustering. Table 3 specifies the best parameters obtained for KNN using grid search algorithm.

Table 3. Parameter tuning for KNN using grid search method.

Parameter	Static MFCC	Dynamic MFCC	LFCC	CC
Neighbors	3	3	3	3

Random Forest (RF) Classifier: This classifier consists of a large number of uncorrelated decision trees that work as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction [3]. It uses the concept of bagging and feature randomness while building the decision trees. Here the classification is done using the concept of majority voting. Table 4 specifies the best parameters obtained for RF using grid search algorithm.

3.3 Evaluation Metric and Procedure

Repeated Stratified K-Fold Valuation: A single run k-fold evaluation can result in a noisy estimation of model performance. The repeated k-fold validation repeated the cross-validation specified number of times which means that instead of increasing the k value to decrease the noise in the evaluation, the number of times the k-fold runs can be increased. The result which we consider is the mean result of all the runs. The term stratified indicates that the proportion of positive and negative classes in the train data and the test data is split equally.

Table 4. Parameter tuning for RF using grid search method.

Parameter	Static MFCC	Dynamic MFCC	LFCC	CC
Maximum depth	20	50	10	50
Samples leaf	1	1	1	1
Estimators	300	150	300	150

Accuracy: Accuracy is a metric that describes how well a model is performing in all the classes. It is used when the dataset is balanced. It is calculated by considering the ratio between TP+TN and TP+TN+FP+FN.

4 Results and Analysis

4.1 Spectrographic Analysis

In Fig. 2, Panel-I and Panel-II represent the spectrographic analysis generated using Librosa [15] for randomly sampled normal and pathological cry signals, respectively. In particular, we took Fourier transform of obtained cepstral features i.e., MFCC, LFCC, and CC. This is indeed a valid representation of log magnitude spectrum as Fourier transform of cepstrum [18]. Figure 2a represents the Static MFCC representations, Fig. 2b represents the dynamic MFCC representations, Fig. 2c represents the LFCC representations, and Fig. 2d represents the cepstral coefficient representations. It can be observed from Fig. 2a and b that there is a difference in the pattern formed by F_0 and its harmonics for normal *versus* pathological cry signals. These differences in the pattern are also visible for LFCC representation as shown in Fig. 2c. However, these differences are more vivid for dynamic MFCC representations as shown in Fig. 2b compared to the static MFCC and LFCC spectrogram. It might be because of the fact that dynamic MFCC can accurately estimate the discriminative acoustic cues of the signal over the entire frequency band considering non-linear aspects of the speech production mechanism and also properties of airflow pattern in the vocal tract system [19]. Furthermore, the results obtained using *10*-fold cross-validation also validate that the dynamic MFCC gives the maximum classification accuracy in this work. On the other hand, the dynamic MFCC is also containing redundant information as seen in the spectrogram compared to the static MFCC, Hence, the average accuracy of static MFCC is greater than that of dynamic MFCC. The features captured by CC are not sufficiently discriminative. Hence, the classifiers are finding it difficult to classify using the features obtained using Cepstral Coefficients.

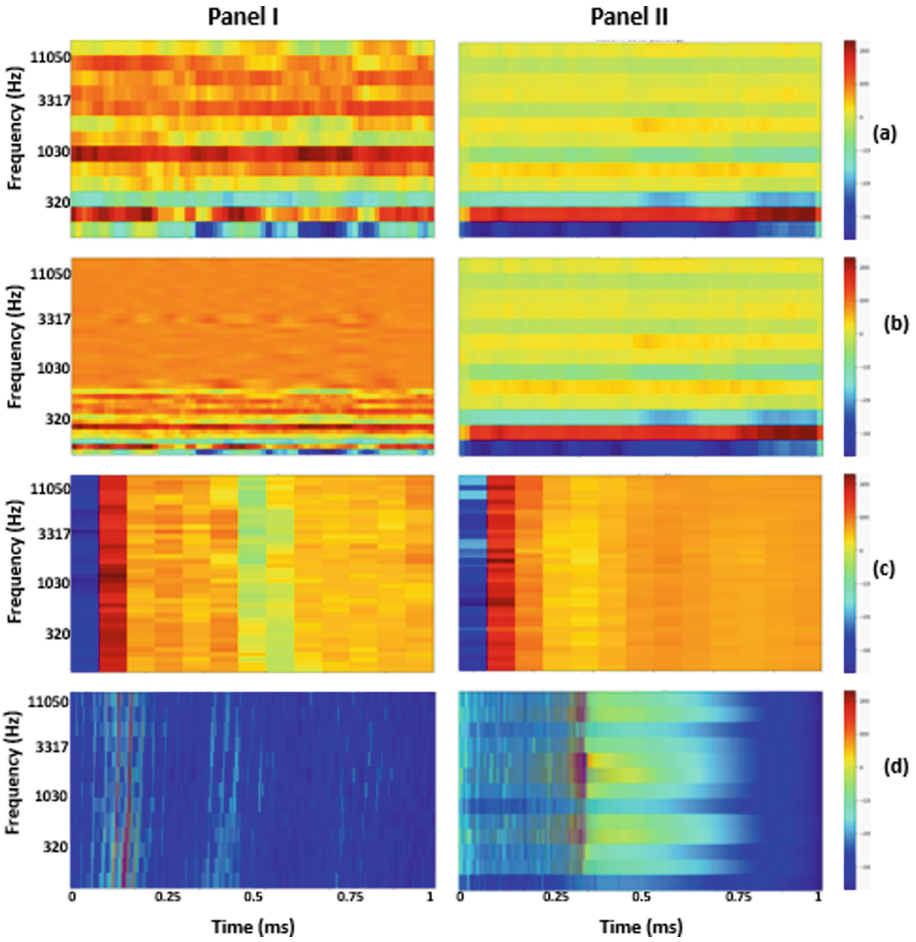

Fig. 2. Panel-I and Panel-II represent the spectrographic analysis (log-magnitude spectrum) of cepstral based features for normal *versus* pathological cry samples, respectively. Figure 2a represents the Static MFCC feature set, Fig. 2b represents the dynamic MFCC representations, Fig. 2c represents the LFCC representations, and Fig. 2d represents the cepstral coefficient representations.

4.2 Performance Evaluation

The performance analysis of various classifiers is done using *3*-repeat *10*-fold stratified cross-validation. The static coefficients of MFCC and dynamic coefficients of MFCC performed similarly resulting in an average fold accuracy across all the classifiers of 95.22 and 93.71%. The features provided by the dynamic MFCC can be redundant features in some cases meaning it degrades the performance of some classifiers, like SVM with a soft margin [4]. In general, the dynamic MFCC represent the trajectory of MFCCs over time by using differen-

tial (delta) and accelerated (delta-delta) coefficients. Hence, even though we are considering 39 coefficients with additional features delta and delta-delta, we are not receiving any additional information other than the information obtained from static MFCC. Hence, the features of the dynamic MFCC feature vector act as *redundant* features reducing the performance of classifiers.

The average repeated *10*-fold accuracy of LFCCs is 94.17%. LFCCs have a linear filterbank meaning that the resolution is better in higher frequencies than the MFCC's higher frequency logarithmic resolution. The LFCC feature vector is having an average accuracy across all the classifiers higher than dynamic MFCC but less than the static MFCC. This indicates that the amount of information obtained from the higher frequencies is significant and cannot be neglected, but at the same time, the average classification accuracy of static MFCC is higher showing that lower frequencies also contain significant crucial information. This comparison also shows the effect of redundancy on the accuracy of dynamic features, when compared with static MFCC and LFCC. The cepstral coefficient feature vector has the worst classification accuracy since it doesn't have any filterbanks as in the case of MFCC and LFCC feature sets.

Coming to the classifiers, all three classifiers handle the redundant data differently as the classification technique of all the 3 classifiers vary. The SVM classifier uses a decision boundary of the linear kernel to classify among the classes. While the KNN classifier uses the clustering concept and assigns a label based on the majority voting of neighbors, the RF classifier assigns a label based on the majority voting on the output of all the decision trees. Hence, we can see a decrease in the performance of some classifiers, when we go from static MFCC to dynamic MFCC feature vector; while the others remain unaffected. The SVM-linear kernel performs better only when the feature extraction is done well else the classification results are poor, and the SVM classifier is affected by the redundant data. The KNN classifier, on the other hand, is a pattern recognition algorithm, which is also another algorithm that is highly dependent on features extracted, it works using the concept of clustering for classification. Hence, when the feature extraction is done properly, the performance of KNN is better than the linear SVM due to the clustering classification technique and the classification with a linear decision boundary is ineffective; which can be observed from the results. The Random Forest classifier tries to outperform both these classifiers (i.e., SVM and KNN) in all the feature extraction techniques. In the RF classifier, the classification accuracy is better compared to the other classifiers across all the feature vectors because the classifier combines many uncorrelated decision trees, as the number of decision trees increases the chances of correct prediction also increase. The RF classifier fails to handle redundant data as the importance score misleads the model.

The average classification accuracy of KNN across all the feature extraction techniques is 89.42%. The number of neighbors parameter for the KNN classifier is obtained using the grid search algorithm and the KNN performed best when the number of neighbors is 3. The average classification accuracy of the Random Forest classifier across all the feature extraction techniques is 90.29%.

The parameters maximum depth, minimum sample leaf, and several estimators are tuned using a grid search algorithm and set to 20, 1, and 300, respectively to obtain the best result of 98.27%. The average classification accuracy of SVM with the linear kernel is 78.82%. The parameter C is tuned using a grid search algorithm and set to 10 for the best result of 87.75%.

From Tables 5, and 6, we can see the effect of window size (20 and 55 ms) on the classification accuracy. The last column of Tables 5 and 6 represents the average performance of various classifiers. The last row of Tables 5 and 6 represents the average accuracy of using various feature extraction techniques. As the window size is increased in any feature extraction technique, the time resolution decreases, and the frequency resolution increases. So, when we increased the window size from 20 to 55 ms, we can see an increase in accuracy across the classifiers indicating that the temporal information can be neglected. The Fig. 3 shows the multi-bar plot of various feature sets for a window size of 55 ms using various classifiers.

Table 5. % Fold accuracy for multiple time-averaged feature vectors with a window size of 20 ms.

Model	Static MFCC	Dynamic MFCC	LFCC	CC	Average acc.
KNN	97.98	98.38	96.74	61.28	88.59
RF	97.66	96.49	96.78	67.98	89.72
SVM linear	86.91	86.99	87.67	58.30	79.96
Average acc.	94.18	93.95	93.73	62.52	

Table 6. % Fold accuracy for multiple time-averaged feature vectors with a window size of 55 ms.

Model	Static MFCC	Dynamic MFCC	LFCC	CC	Average acc.
KNN	98.42	**98.48**	97.50	63.30	89.42
RF	97.92	96.61	97.27	69.37	90.29
SVM linear	85.44	86.07	87.75	56.05	78.82
Average acc.	93.92	93.72	94.17	62.90	

Since temporal information is not very important in the classification of normal *versus* pathological cry, we averaged the temporal-axis of the matrix obtained from the feature extraction technique and converted it into a *1*-D vector. The results show that there is not much loss of information as the maximum stratified *10*-fold accuracy obtained is 98.48% (Table 6). This also reduces the computational complexity while feeding the features into the classifiers or deep learning architectures.

The secondary goal is to keep the false positive count to a minimum so that the misclassification of pathology cry as normal is less which is very important in

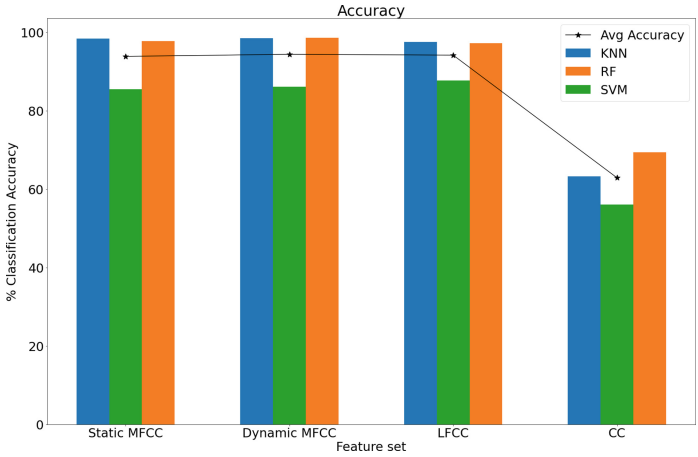

Fig. 3. % Classification accuracy of various feature vectors of window size 55 ms with various classifiers.

realistic scenarios. This is consistently achieved in static MFCC feature vectors across all three classifiers and the best results are seen using dynamic MFCC feature extraction with KNN and RF as classifiers as shown in Fig. 4 and Table 7.

Table 7. Confusion Matrix for Dynamic MFCC using various classifiers.

Classifier	Class	Normal	Pathology
KNN	Normal	**300**	4
	Pathology	5	**370**
RF	Normal	**290**	15
	Pathology	2	**380**
SVM	Normal	**260**	43
	Pathology	61	**320**

The results obtained apparently gave counterintuitive analysis, i.e., dynamic MFCC performed better than static MFCC for a few classifiers. This might be because of the fact that each classifier selected has a different method of deciding the classification boundary, hence the additional data obtained using dynamic MFCC affects the classifiers uniquely. However, it should be noted that the gradient time-averaged feature vector implicitly captures dynamic information, as it is the concatenation of static features, delta features, and delta-delta features.

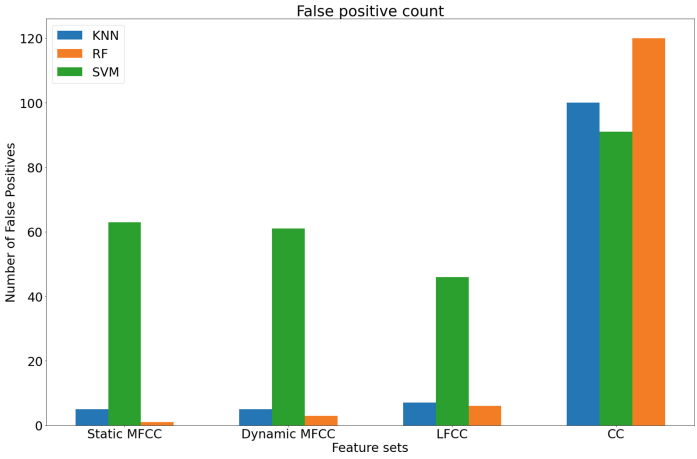

Fig. 4. False positive count of various feature vectors of window size 55 ms with various classifiers.

5 Summary and Conclusion

In this work, we presented a comparative study of various time-averaged feature extraction techniques such as MFCC, LFCC, and Cepstral Coefficients. The effect of different window sizes was also studied in this work. It was found that the dynamic MFCC feature vector with a window size of 55 ms along with the KNN classifier results in the best relative classification accuracy of 98.48% with 5 false positives. It was also observed that there was an increase in the classification accuracy of about 0.5% on KNN and RF classifiers with static MFCC, however, there was a minor change in the classification accuracy by using dynamic MFCC. Hence, it can be concluded that the infant's cries contain discriminative cues in the spatial or the frequency plane rather than the temporal or time plane. Hence, the feature extraction techniques are averaged across the time-axis reducing a 2-D array into a 1-D array for each audio file. It was also observed that the amount of information in lower frequencies is slightly higher than in the higher frequencies. However, the linear kernel SVM failed to perform well enough compared to the other classifiers. To that effect, our future work will be directed toward exploring non-linear kernels, such as Radial Basis Functions (RBF), and polynomial kernels for SVM. Furthermore, Deep learning architectures like Convolutional Neural Network (CNN) and Light CNN (LCNN) along with data augmentation can be explored for the classification task across various feature extraction techniques.

Acknowledgments. The authors sincerely thank the organizers the National Institute of Astrophysics and Optical Electronics, CONACYT Mexico for the statistically meaningful Baby Chilanto database, the Ministry of Electronics and Information Technology (MeitY), New Delhi, Govt. of India, for sponsoring a consortium project titled 'Speech Technologies in Indian Languages' under 'National Language Translation

Mission (NLTM): BHASHINI', subtitled 'Building Assistive Speech Technologies for the Challenged' (Grant ID: 11(1)2022-HCC (TDIL)). We also thank the consortium leaders Prof. Hema A. Murthy and Prof. S. Umesh of IIT Madras, and the authorities of DA-IICT Gandhinagar, India for their support and cooperation to carry out this research work.

References

1. Alaie, H.F., Abou-Abbas, L., Tadj, C.: Cry-based infant pathology classification using GMMs. Speech Commun. **77**, 28–52 (2016)
2. Bakshi, A., Kopparapu, S.K., Pawar, S., Nema, S.: Novel windowing technique of MFCC for speaker identification with modified polynomial classifiers. In: 2014 5th International Conference—Confluence The Next Generation Information Technology Summit (Confluence), pp. 292–297 (2014). https://doi.org/10.1109/CONFLUENCE.2014.6949342, Accessed: 15 Aug 2022
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006)
4. Cohen, G., Hilario, M., Sax, H., Hugonnet, S., Geissbuhler, A.: Learning from imbalanced data in surveillance of nosocomial infection. Artif. Intell. Med. **37**(1), 7–18 (2006)
5. Cover, T.M.: Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electron. Comput. **3**, 326–334 (1965)
6. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. **28**(4), 357–366 (1980)
7. Dewi, S.P., Prasasti, A.L., Irawan, B.: Analysis of LFCC feature extraction in baby crying classification using KNN. In: 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), pp. 86–91. IEEE (2019)
8. Dewi, S.P., Prasasti, A.L., Irawan, B.: The study of baby crying analysis using MFCC and LFCC in different classification methods. In: 2019 IEEE International Conference on Signals and Systems (ICSigSys), pp. 18–23. IEEE (2019)
9. Engelsma, J.J., Deb, D., Cao, K., Bhatnagar, A., Sudhish, P.S., Jain, A.K.: Infant-id: fingerprints for global good. IEEE Trans. Pattern Anal. Mach. Intell. **44**, 3543–3559 (2021)
10. Fairbanks, G., Wiley, J.H., Lassman, F.M.: An acoustical study of vocal pitch in seven- and eight-year-old boys. Child Dev. **20**(2), 63–69 (1949). https://www.jstor.org/stable/1125607. Accessed 15 Aug 2022
11. 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, vol. 4, pp. 3140–3145. IEEE (2003)
12. Hossan, M.A., Memon, S., Gregory, M.A.: A novel approach for MFCC feature extraction. In: 2010 4th International Conference on Signal Processing and Communication Systems, pp. 1–5 (2010). https://doi.org/10.1109/ICSPCS.2010.5709752. Accessed 11 Aug 2022
13. Irwin, O.C., Curry, T.: Vowel elements in the crying vocalization of infants under ten days of age. Child Dev. **12**(2), 99–109 (1941). https://www.jstor.org/stable/1125343. Accessed 12 Aug 2022

14. Ji, C., Mudiyanselage, T.B., Gao, Y., Pan, Y.: A review of infant cry analysis and classification. EURASIP J. Audio, Speech, Music. Process. **2021**(1), 1–17 (2021). https://doi.org/10.1186/s13636-021-00197-5
15. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference, vol. 8, pp. 18–25. Citeseer (2015)
16. Osmani, A., Hamidi, M., Chibani, A.: Platform for assessment and monitoring of infant comfort. In: AAAI Fall Symposium Series, p. 2017. Virginia, Arlington (2017)
17. Patil, H.A.: Infant identification from their cry. In: 2009 Seventh International Conference on Advances in Pattern Recognition, pp. 107–110. IEEE (2009)
18. Patil, H.A.: Cry baby: using spectrographic analysis to assess neonatal health status from an infant's cry. In: Newtein, A. (ed.) Advances in Speech Recognition, pp. 323–348. Springer, Berlin (2010)
19. Quatieri, T.F.: Discrete-Time Speech Signal Processing: Principles and Practice. 1st edn, Pearson Education India (2015)
20. Rosales-Pérez, A., Reyes-García, C.A., Gonzalez, J.A., Arch-Tirado, E.: Infant cry classification using genetic selection of a fuzzy model. In: Iberoamerican Congress on Pattern Recognition, pp. 212–219. Springer, Berlin (2012)
21. Rosales-Pérez, A., Reyes-García, C.A., Gonzalez, J.A., Reyes-Galaviz, O.F., Escalante, H.J., Orlandi, S.: Classifying infant cry patterns by the genetic selection of a fuzzy model. Biomed. Signal Process. Control **17**, 38–46 (2015)
22. Wasz-Höckert, O., Partanen, T., Vuorenkoski, V., Michelsson, K., Valanne, E.: The identification of some specific meanings in infant vocalization. Experientia **20**(3), 154–154 (1964)
23. Wasz-Hockert, O., Valanne, E., Vuorenkoski, V., Michelsson, K., Sovijarvi, A.: Analysis of some types of vocalization in the newborn and in early infancy. In: Annales Paediatriae Fenniae, vol. 9, pp. 1–10 (1963)
24. Xie, Q., Ward, R.K., Laszlo, C.A.: Automatic assessment of infants' levels-of-distress from the cry signals. IEEE Trans. Speech Audio Process. **4**(4), 253 (1996)