CLASSIFICATION OF INFANT CRIES USING SOURCE, SYSTEM AND SUPRA-SEGMENTAL FEATURES

Avinash Kumar Singh, Jayanta Mukhopadhyay and K. Sreenivasa Rao

School of Information Technology
Indian Institute of Technology Kharagpur, Kharagpur - 721302, West Bengal, India.
E-mail: avinashshalini11@gmail.com, jay@cse.iitkgp.ernet.in, ksrao@iitkgp.ac.in

ABSTRACT

In this paper, source, system and supra-segmental features are explored for recognition of infant cry. Different types of infant cries considered in this work are hunger, pain and wet-diaper. In this work, mel-frequency cepstral coefficients (MFCC), residual MFCC (RMFCC), implicit LP residual features, features from modulation spectrum and time domain envelope features are used for representing the infant cry specific information from the acoustic signal. Gaussian Mixture Models (GMM) are used for classifying the above mentioned cries from the features proposed in this work. GMM models are developed separately by using the proposed features. Infant cry database collected under telemedicine project (eNPCS) at IIT-KGP has been used for carrying out this study. The recognition performance of the developed GMM models is observed to be varying significantly based on the features. Results have indicated that, the proposed features have complementary evidences in view of discriminating the infant cries. For enhancing the recognition performance, GMM models developed using various features are combined using score level fusion. The recognition performance using combination of evidences is found to be superior over individual systems.

Index Terms— Infant cry recognition, Gaussian Mixture Model, Spectral features, Modulation Spectrum, Time domain envelope.

1. INTRODUCTION

It is important to analyze the new born infant cry signals to aid clinical diagnosis. Infants have only one way communication with others by crying. An infant's cry contains a lot of information about the baby, feeling hunger, pain, sleepiness, uncomfortable due to wet-diaper etc. In general, experienced mothers can normally recognize their baby needs by differentiating between their types of cries. But there should be an automatic way to analyze the meaning of the cries, so that their needs can be understood. A new born infant cry is characterized by very high fundamental frequency (F0) with sudden changes and voiced or unvoiced features of very short duration. The highly varying vocal tract resonance frequencies need to be tracked accurately. However, there is a need for pre-processing to reduce the variability before analysis. There are number of ways to analyze the infant cries. Most of the state-of-art literatures used spectral features such as either Linear Prediction Ceptstral Coefficients (LPCC) or Mel Frequency Cepstral Coefficients (MFCC) to discriminate the cries [1, 2, 3]. In [1], Linear Prediction Coefficients (LPC) and intensity features are used to discriminate cries due to pain and hunger using neural networks. In [2], MFCC and LPC features were used to discriminate normal vs pathological cry. In [3] MFCC and LPCC are explored for discriminating pain vs non-pain. Ramu Reddy et al., have explored spectral and prosodic features for characterizing the infant cries [4]. In their work, mel-frequency cepstral coefficients (MFCC) are used to represent the spectral information, and short-time frame energies (STE) and pause duration are used for representing the prosodic information. They have considered three cries namely, hunger, wet-diaper and pain. Support Vector Machines (SVM) are used to capture the discriminative information with respect to cries from the spectral and prosodic features.

It is observed, that most of the researchers discriminated the cries as bi-classification problem such as pain vs no-pain, pain vs hunger and normal vs disease cries. In most of the works only spectral features such as MFCC and LPCC are used for classifying the cries. In literature combination of spectral, source and prosodic features are explored for improving the performance of various speech tasks [5, 6, 7, 8, 9, 10, 11]. Therefore, in this work we have explored residual MFCC (RMFCC), implicit LP residual features, features from modulation spectrum and time domain envelope features, in addition to conventional spectral features. In above features, residual MFCC and implicit LP residual features represent excitation source information, modulation spectrum and time domain envelope features represent suprasegmental information, and MFCC features represent spectral or vocal tract system characteristics. In this study, we considered three cries namely hunger, pain and wet-diaper. In this paper, we have considered only MFCCs for representing the system or spec-

978-1-4673-5842-2/13/$31.00 © 2013 IEEE

tral characteristics, due to their better performance compared to LPCCs [2] [3] in discriminating the cries. Gaussian Mixture Model (GMM) are used for developing the models to capture the distribution of features specific to each cry.

Rest of the paper is organized as follows: Following section describes the database used for infant cry discrimination. Section 3 describes the proposed features for discriminating the infant cries. Development of GMM models using proposed features is discussed in Section 4. Recognition performance of the developed GMM models with proposed features is explained in Section 5. Summary of this paper and the future work needed to improve the recognition performance are described in the final section.

2. INFANT CRY DATABASE

The infant cry speech corpus used for this study is collected from 120 infants in Neonatal Intensive Care Unit (NICU) hospital (located in Kolkata) which is specialized in the care of ill or premature newborn infants. The database consists of three different types of cries namely wet-diaper, hunger and pain. The total number of cry clips collected for wet-diaper, hungry and pain are 30, 60 and 30 respectively. The duration of clips for wet-diaper, hungry and pain cries varies from 10-49, 12-55 and 9-40 sec respectively. The total duration of the collected cry clips for wet-diaper, hungry and pain are 937 sec, 1874 sec, and 725 sec respectively. The infants selected for recording the cries are within the age group of 12-40 weeks old. Recording has been done by using a Sony digital recorder with sampling rate of 44.1 kHz. For this study the recorded data is down sampled to 16 kHz and represented each sample as 16 bit number.

3. FEATURES

In this work, we have proposed implicit LP residual features, residual MFCCs, MFCCs, modulation spectrum features and time domain envelope features for classifying the infant cries. The following subsections discuss the details of extraction of the proposed features.

3.1. Implicit LP residual features

In general, LP residual signal is considered as excitation source component of speech signal. It is derived from speech signal by inverse LP filter formulation. The inverse LP filter will suppress or remove the vocal tract information, and hence its outcome may be viewed as excitation source information. The LP residual signal looks like a noise signal due to removal of 1st and 2nd order correlations by LP inverse filter. For capturing the higher order nonlinear relations present in LP residual, the sequence of LP residual samples at different levels are used as feature vectors. In this study, we have explored the sequence of LP residual samples within 5 ms, 20 ms and 200 ms speech segments for representing the implicit excitation source information at sub-segmental, segmental and supra-segmental levels, respectively. The detailed explanation about these features is given in [12].

3.2. Residual Mel-Frequency Ceptstral Coefficients (RMFCC)

Residual MFCCs represent the spectral characteristics of the excitation source signal. The basic steps in determining the RMFCCs are similar to conventional MFCCs. The details of MFCC feature extraction are given in the following subsection.

3.3. Mel-Frequency Ceptstral Coefficients (MFCC)

The characteristics of the cry can be attributed to the characteristics of vocal tract system, excitation source, and suprasegmental characteristics. At the segmental level, cry specific vocal tract information is mainly represented by spectral features. The spectral peaks, bandwidths and slopes are unique for each cry. Hence, if we capture this information one can discriminate the cries. In this work, we exploited MFCC features to capture the discriminated spectral information of each cry. MFCCs are determined from cry signal using the following steps.

1. Pre-emphasize the infant cry signal.
2. Divide the cry signal into sequence of frames with a frame size of 20 ms and a shift of 10 ms. Apply the Hamming window over each of the frames.
3. Compute magnitude spectrum for each windowed frame by applying DFT.
4. Mel spectrum is computed by passing the DFT signal through mel filter bank.
5. DCT is applied to the log mel frequency coefficients (log mel spectrum) to derive the desired MFCCs.

In this study, we used 13 dimensional spectral features representing 13 MFCCs. The MFCCs are derived from a speech frame of 20 ms with a frame shift of 10 ms. For deriving the MFCCs, 24 filter bands are used.

3.4. Modulation Spectrum Features

In an acoustic signal, it is assumed that information-bearing components are of the form $x(t) = m(t).c(t)$

where $x(t)$ is an observed time-domain signal, and $m(t)$ is a low-frequency modulator multiplied by a high-frequency carrier $c(t)$. In analysis of modulation spectra involves demodulating the speech signal and estimating discrete time $m(n)$ and $c(n)$ from $x(n)$. After estimating $m(n)$ represents time domain envelope fluctuation that can be analysed

by Fourier transform for modulation analysis and filtering methods. The modulated signal is given by

$$x(n) = \sum_{k=0}^{K-1} m_k(k).c_k(n)$$

where $x(n)$ is an observed discrete-time full band signal, $m_k(n)$ and $c_k(n)$ are the respective k^{th} modulator and carrier waveforms, the (.) operator denotes sample by sample multiplication, and K is a finite number. Using spectral centre of gravity method spectral concentration can be tracked within a sub-band over time as an estimate of carrier frequency. Thus modulation and carrier frequency are separated. The modulation spectrum is the windowed Fourier transform across time

$$P_k(i) = \sum_{k=0}^{K-1} w(n).m_k(k)e^{-j\frac{2\pi i}{N}n}$$

3.5. Time domain envelope features

All pole estimation for speech signals is obtained by establishing a band limited interpolation of the observed spectrum by using a true envelope estimator. It is evident fact that a speaker specific information is contained in the inter band correlations of narrow band temporal envelopes. Analysis of these temporal envelope features may be useful in discriminating the cries in ambiguous situations. In this study, discrete cosine transform coefficients derived from the envelope signal are used for representing the feature vectors. This obtained feature is more indicative of modulation index. It can be hypothesized that all pole envelope of signal gives unique feature for each cry.

4. GAUSSIAN MIXTURE MODEL

The Gaussian Mixture Model is one of the statistically mature method for unsupervised clustering [13]. The number of Gaussians in the mixture model are also known as components. Comparing with the k-means clustering, one component can be thought as one cluster's information. So, the number of components indicate the number of clusters in which data points are to be distributed in order to cover the local variations.

The complete Gaussian Mixture Model is parametrized by the mean vector, diagonal of covariance matrix and the mixture weight of each component. These parameters are collectively represented by the following notation .

$$\lambda = \{\rho_i, \vec{\mu_i}, \Sigma_i\} \ , \quad i = 1, 2, ..., M \tag{1}$$

where, M we have chosen as 64. ρ_i is the weight, $\vec{\mu_i}$ is the mean vector and Σ_i is the covariance matrix of i^{th} component. The mixture weights satisfy the constraint that, $\Sigma_{i=1}^{M}\rho_i = 1$. These parameters are initialized using $k-$means clustering on the training set of feature vectors, with $k = M$.

A well known Expectation Maximization (EM) algorithm is used for finding the maximum likelihood estimates of the parameters. EM is an iterative method that alternates between performing an expectation (E) step, which computes an expectation of the log likelihood with respect to the current estimate of distribution and a maximization (M) step which computes the parameters that maximize the expected likelihood found in the E step. These parameters are then used for E step of next iteration. We perform 100 iterations of EM steps in our experiment. Once the parameters are re-estimated by EM algorithm, the training phase is completed. Now, given a test vector $\vec{x}$ of dimension D, the score is generated by the equation 2.

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} \rho_i b_i(\vec{x}) \tag{2}$$

where, $b_i(\vec{x})$ is the component density of i^{th} component, given by the equation 3

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} \ |\Sigma_i|^{\frac{1}{2}}} \ exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu_i})' \Sigma_i^{-1} (\vec{x} - \vec{\mu_i})\right\} \tag{3}$$

The GMM can have several different forms depending on the choice of covariance matrices. The model can have one covariance matrix per Gaussian component or one covariance matrix for all Gaussian components in a cry model or a single covariance matrix shared by all cry models. The covariance matrix can be full or diagonal. Here, diagonal matrices are primarily used. The choice of model configuration (number of components, full or diagonal covariance matrices and mixture weights) is often determined by the amount of data available for estimating the GMM parameters and how the GMM is used in a particular application.

In this work, around 75% of the data is used for developing GMM models, and the remaining 25% data is used for validation. For each cry category, a separate GMM is developed to represent the distribution of feature vectors belongs to that particular cry. In this work, the infant-cry recognition system consists of 3 GMMs followed by a decision device. Separate recognition systems are developed for analyzing the discriminating capability of each of the proposed features. For enhancing the recognition performance, evidences from the individual systems are combined with appropriate weights. In this work, seven Infant Cry Recognition (ICR) systems are developed using individual proposed features.

1. ICRS-1: Infant cry recognition system developed using sub-segmental implicit LP residual features.

2. ICRS-2: Infant cry recognition system developed using segmental implicit LP residual features.

3. ICRS-3: Infant cry recognition system developed using supra-segmental implicit LP residual features.

4. ICRS-4: Infant cry recognition system developed using RMFCC features.

5. ICRS-5: Infant cry recognition system developed using MFCC features.

6. ICRS-6: Infant cry recognition system developed using modulation spectrum features.

7. ICRS-7: Infant cry recognition system developed using time domain envelope features.

Given training vectors and a GMM configuration, we wish to estimate the parameters of the GMM, which in some sense best matches the distribution of the training feature vectors. There are several techniques available for estimating the parameters of a GMM . By far the most popular and well established method is maximum likelihood (ML) estimation.

5. RESULTS AND DISCUSSION

In this work, we have analyzed the recognition performance using 2, 5 and 10 sec cry samples. Recognition performance is observed to be improved by using 5 and 10 sec samples compared to 2 secs samples. Among 5 and 10 sec samples, not much variation in recognition performance is observed. The results mentioned in this paper are confined to 5 sec cry samples. In this study with 5 secs duration, the number of test samples correspond to hunger, wet-diaper and pain are 125, 75 and 50 respectively. The performance of the ICR system using proposed features is represented in the form of confusion matrix. The diagonal elements of the confusion matrix represents the correct classification performance of cries. Other than the diagonal elements indicates the misclassification performance. The details of the recognition performance using proposed features is discussed in following subsections.

5.1. Implicit LP residual features

The recognition performance of the ICRS-1 developed using Sub-segmental implicit LP residual features is shown in Table 1. Average recognition performance is observed to be about 35%. Recognition performance of ICRS-2 and ICRS-3 developed using segmental and suprasegmental implicit LP residual features is given in Tables 2 and 3, respectively. Average recognition performance using segmental and suprasegemntal implicit LP residual features is observed to be 54% and 49%, respectively. From the results, it is observed that recognition performance from the above 3 systems (ICRS-1, ICRS-2 and ICRS-3) is complementary to some extent. Therefore, we have explored the combination of their evidences at score level for improving the recognition performance further. Table 4 shows the recognition performance of combined system developed by score level fusion of individual systems. The recognition performance is found to be improved compared to individual systems. Average recognition performance of the combined system developed using implicit LP residual features extracted from sub-segmental, segmental and suprasegmental levels is found to be 59%.

Table 1. Performance of the infant cry recognition system developed by using sub-segmental implicit LP residual features.

	Recognition performance (%)		
	Hunger	Wet-diaper	Pain
Hunger	30.00	32.86	37.14
Wet-diaper	19.35	48.39	32.26
Pain	43.64	29.09	27.27

Table 2. Performance of the infant cry recognition system developed by using segmental implicit LP residual features.

	Recognition performance (%)		
	Hunger	Wet-diaper	Pain
Hunger	62.07	17.24	20.69
Wet-diaper	20.98	58.06	20.96
Pain	45.45	14.55	40.00

Table 3. Performance of the infant cry recognition system developed by using suprasegmental implicit LP residual features.

	Recognition performance (%)		
	Hunger	Wet-diaper	Pain
Hunger	60.71	30.00	9.29
Wet-diaper	23.07	73.85	3.08
Pain	47.27	41.82	10.91

Table 4. Performance of the infant cry recognition system developed by combining the evidences from subsegmental, segmental and suprasegmental implicit LP residual features.

	Recognition performance (%)		
	Hunger	Wet-diaper	Pain
Hunger	61.43	28.57	10.0
Wet-diaper	30.13	59.67	10.20
Pain	34.97	11.31	53.72

5.2. Residual MFCC features

The recognition performance of the ICRS-4 developed using MFCC features extracted from LP residual signal is shown in Table 5. The average recognition performance is found to be 64%. These features discriminated all three cries in an uniform manner to some extent. Pain was recognised with 50% accuracy, whereas hunger and wet-diaper are recognised with 65% and 67% accuracy, respectively.

5.3. MFCC features

The recognition performance of the ICRS-5 developed using MFCC features extracted from cry signal is shown in Table

Table 5. Performance of the infant cry recognition system developed by using MFCC features extracted from LP residual signal (RMFCC).

	Recognition performance (%)		
	Hunger	Wet-diaper	Pain
Hunger	64.90	24.07	11.03
Wet-diaper	23.46	66.73	9.81
Pain	47.21	2.71	50.08

6. The average recognition performance is found to be 63%. These features also discriminated all three cries in an uniform manner similar to RMFCCs. Pain was recognised with 47% accuracy, whereas hunger and wet-diaper are recognised with 61% and 69% accuracy, respectively.

Table 6. Performance of the infant cry recognition system developed by using MFCC features extracted from cry signal (MFCCs).

	Recognition performance (%)		
	Hunger	Wet-diaper	Pain
Hunger	60.82	29.05	10.13
Wet-diaper	21.07	68.73	10.20
Pain	50.62	2.51	46.87

5.4. Modulation spectrum features

The recognition performance of the ICRS-6 developed by using modulation spectrum features extracted from cry signal is shown in Table 7. The average recognition performance is found to be 43%. Among three cries, accuracy of recognition of pain is very low. From the results, it is observed that this system is slightly biased towards hunger. Hence, about 50% of the test samples in each cry category are classified as hunger. This may be due to presence of large number of hunger cry samples compared to others.

Table 7. Performance of the infant cry recognition system developed by using modulation spectrum features extracted from cry signal.

	Recognition performance (%)		
	Hunger	Wet-diaper	Pain
Hunger	59.64	31.22	9.14
Wet-diaper	48.01	46.92	5.07
Pain	58.35	21.52	20.13

5.5. Time domain envelope features

The recognition performance of the ICRS-7 developed by using time domain envelope features is shown in Table 8. The average recognition performance is found to be 64%. These features have recognised the pain with very high accuracy of about 81%, compared to other proposed features.

Table 8. Performance of the infant cry recognition system developed by using time domain envelope features.

	Recognition performance (%)		
	Hunger	Wet-diaper	Pain
Hunger	52.15	24.32	23.53
Wet-diaper	21.19	61.73	17.08
Pain	13.56	5.23	81.21

5.6. Combination of ICR systems using score level fusion

The complementary evidences offered by various ICR systems developed using individual features may be exploited by combining their evidences using score level fusion. Table 9 shows the recognition performance by combining the evidences from individual systems. From the results, it is observed that the recognition performance has been enhanced by combining the evidences from the individual systems. This may be due to complementary evidences from the individual systems. The average recognition performance of the system developed by combining the evidences from all the proposed features is about 68%.

6. SUMMARY AND CONCLUSIONS

In this work, source features represented by implicit LP residual features and residual MFCCs, system features represented by MFCC features and supra segmental features represented by modulation spectrum features and time domain envelope features are explored for classifying the infant cries. GMMs were used as classification models for developing different ICR systems. The performance of the ICR systems developed by implicit LP residual features, residual MFCCs, MFCCs, Modulation spectrum features and time domain envelope features are observed to be 59%, 64%, 63%, 43% and 64%, respectively. Fusion techniques were explored by combining the evidences of individual ICR systems at the score level. The performance of the ICR system was improved by score level fusion of evidences from the individual systems. The overall recognition performance of the combined system using score level fusion is found to be 68%. The performance of ICR system may be improved by combining the prosodic features to the proposed features. In this work, ICR systems are developed using GMM models, which are basically generative models. The performance may be enhanced by

Table 9. Performance of the infant cry recognition system developed by combining the evidences from individual systems.

ICR Systems	Recognition performance (%)
Sub-segmental implicit LP residual features	35
Segmental implicit LP residual features	54
Supra-segmental implicit LP residual features	49
Implicit LP residual features (Sub-segmental + Segmental +Supra-segmental)	59
Residual MFCC features (RMFCCs)	64
MFCC features	63
RMFCCs + MFCCs	64
Modulation spectrum features	43
Time domain envelope features	64
Implicit LP residual + RMFCCs + MFCCs + Modulation spectrum features Time domain envelope features	68

exploring the hybrid models consists of both generative and discriminative models.

Acknowledgements

The work presented in this paper was performed at IIT Kharagpur as a part of the project "Development of a Web enabled e-Healthcare System for Neonatal Patient Care Services (eNPCS)" sponsored by Ministry of Communication and Information Technology (MCIT), Government of India. Our special thanks to the project team at IIT, Kharagpur and doctors of Department of Neonatology, SSKM Hospital for collecting the cry samples of infants.

7. REFERENCES

[1] O. O. Garca and C. A. R. Garca, "Applying Scaled Conjugate Gradient for the Classification of Infant Cry with Neural Networks," in *European Symposium on Artificial Neural Networks*, (Bruges, Belgium), pp. 349–354, Apr 2003.

[2] 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," (St. Petersburg, Russia), Sept. 2004.

[3] Y. Abdulaziz and S. M. S. Ahmad, "Infant Cry Recognition System: A Comparison of System Performance based on Mel Frequency and Linear Prediction Cepstral Coefficients," in *IEEE*, (Shah Alam,Selangor,Malaysia), pp. 260–263, March 2010.

[4] R. R. Vempada, S. A. K. B, and K. S. Rao, "Characterization of Infant Cries using Spectral and Prosodic features," in *National Conference on Communications (NCC-2012)*, (IIT Kharagpur, Kharagpur, India.), Feb. 2012.

[5] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 972–980, May 2006.

[6] K. S. Rao, *Acquisition and incorporation prosody knowledge for speech systems in Indian languages*. PhD thesis, Dept. of Computer Science and Engineering, Indian Institute of Technology Madras, Chennai, India, May 2005.

[7] K. S. Rao and B. Yegnanarayana, "Modeling durations of syllables using neural networks," *Computer Speech and Language*, vol. 21, pp. 282–295, Apr. 2007.

[8] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, (Orlando, Florida, USA), pp. 541–544, May 2002.

[9] L. Mary, K. S. Rao, S. V. Gangashetty, and B. Yegnanarayana, "Neural network models for capturing duration and intonation knowledge for language and speaker identification," in *Int. Conf. Cognitive and Neural Systems*, (Boston, MA, USA), May 2004.

[10] K. S. Rao and B. Yegnanarayana, "Intonation modeling for Indian languages," in *Proc. Int. Conf. Spoken Language Processing*, (Jeju Island, Korea), pp. 733–736, Oct. 2004.

[11] K. S. Rao, "Voice conversion by mapping the speaker-specific features using pitch synchronous approach," *Computer Speech and Language*, vol. 23, no. 2, pp. 240–256, 2009.

[12] D. Pati and S. R. M. Prasanna, "Subsegmental, segmental and suprasegmental processing of linear prediction residual for speaker information," *International Journal of Speech Technology, (Springer)*, vol. 14, pp. 49–63, Feb. 2011.

[13] D. A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," *Speech Communication*, vol. 17, no. 2, pp. 91–108, 1995.