Infant Cry Signal Detection, Pattern Extraction and Recognition

Lichuan Liu, Yang Li, Kevin Kuo
Department of Electrical Engineering
Northern Illinois University
DeKalb, Il 60115, USA
e-mail: liu@niu.edu, yanglee@yahoo.com, Kevinkuo@gmail.com

***Abstract*—The cry signals generated by infants serves as the primary communication for infants. Cry signals can provide insight into their wellbeing. This paper proposes to use the speech signal identification technique to recognize infant cry signals. Advanced signal processing methods are used to analyze the infant cry by using audio features in the time and frequency domains in an attempt to classify each cry to a specific need. The features extracted from audio feature space include linear predictive coding (LPC), linear predictive cepstral coefficients (LPCC), Bark frequency cepstral coefficients (BFCC) and Mel frequency cepstral coefficients (MFCC). The primary classification technique used were: nearest neighbor approach, neural networks method. The cry recognition of specific infants yielded promising results.**

Keywords-infant cry; cry detection; pattern extraction; recognition; nearest neighborhood; neural network

I. INTRODUCTION

It is well know that the cry serves as the primary means of communication for very young age infants. It is possible for experts to distinguish infant cries based on training and experience. However, it is difficult for new parents and for inexperienced child care givers to interpret infant cries. Since the infant cry can be distinguished by experts, it is possible to extract audio features from the infant cry such that it can be uniquely differentiated from cries of a different reason [1], [2].

Prior works on infant cry analysis have either investigated the difference between normal and pathological (deaf or hearing loss infants) cries [1], [2], or they have attempted to differentiate conditional cries [3] such as pain from shots, fear from jack-in-the box toys, and frustration from body restraints. Cry Translator, a commercially available product, claims that it can identify five distinct cries: hunger, sleep, discomfort, stress and boredom. Their results were achieved through a universal algorithm that takes into account the pattern of loudness, pitch, tone and inflection of the infant cry. [3]. We have used the speech recognition method to recognize the individual infant cry signals and found that the classification algorithm can recognize infant cry signal with 70% classification rate [4], [5]. In this paper, we try to extend the cry signal recognition algorithm from the individual infant recognition to infant-independent recognition.

Analyzing the infant cry signals can provide a non-invasive diagnostic of the condition of the infant [6]. Using cry as a diagnostic tool plays an important role in the following situations: medical problems in which there is currently no diagnostic tool available, for example SIDS, problems in developmental outcome and colic; medical problems in which early detection is possible only by invasive procedures; medical problems which may be readily identified but would benefit from an improved ability to define prognosis [7].

II. CRY SIGNAL DETEDCTION

In order to accurately detect potential periods of cry activity, two short term signal detection techniques are used.

A. Short-Time Energy

Short-time energy (STE) is defined as the average of the square of the sample values in a suitable window. It can be mathematically described as follows [9]:

$$STE(n) = \frac{1}{N}\sum_{m=0}^{N-1}[w(m)x(n-m)]^2 \qquad (1)$$

where $w(m)$ are coefficients of a suitable window function of length N. The Hamming window was chosen as it minimizes the maximum side lobe in the frequency domain.

A For our signals of 8 kHz sampling frequency, a window of 128 samples (~16 ms) was used. STE estimation is useful as a speech detector because there is a noticeable difference between the average energy between voiced and unvoiced speech, and between speech and silence [8]. This technique is usually paired with short-time zero crossing for a robust detection scheme.

B. Short-Time Zero Crossing

Short-time zero crossing (STZC) is defined as the rate at which the signal changes sign. It can be mathematically described as follows [10]:

$$Z(n) = \frac{1}{N}\sum_{m=0}^{N-1}|\,\mathrm{sign}(x(n-m)) - \mathrm{sign}(x(n-m-1))\,| \qquad (2)$$

where $\mathrm{sign}(x(m)) = \begin{cases} 1 & x(m) \geq 0 \\ -1 & x(m) < 0 \end{cases}$

STZC estimation is useful as a speech detector because

there are noticeable fewer zero crossings in voiced speech as compared with unvoiced speech.

III. Feature Extraction

Audio feature extractions hinges upon using techniques in digital signal processing of audio signals to quantize acoustic information in a manner that makes classification tractable. This process will be discussed for use in the time and frequency domain in the following sections.

A. Linear Predictive Coding

The waveforms of two similar sounds will also be similar. If two infant cries have very similar waveforms, it stands to reason that they should possess the same impetus. However, it is impractical to conduct a sample by sample full comparison between cry signals due to the computational complexity. In order to improve the solution of the time domain comparison of infant cry signals, linear predictive coding (LPC) is applied.

The linear predictive coding (LPC) algorithm produces a vector of coefficients that represent a spectral shaping filter [9], [10]. This shaping filter is an all-pole filter represented as [11]:

$$H(z) = \frac{1}{1 - \sum_{i=1}^{M} a_i z^{-i}} \tag{3}$$

where a_i are the linear prediction coefficients and M is the number of poles (the roots of the denominators in the z transform). Equation (3) can describe the present sample of the speech as a linear combination of the past M samples of the speech such that:

$$\hat{x}(n) = \sum_{i=1}^{M} a_i x(n-i) \tag{4}$$

where $\hat{x}(n)$ is the predicted value of $x(n)$, $\{a_i\}$ are the linear prediction coefficients and M is the number of of the all-pole filter. Then the coefficients$\{a_i\}$ can be estimated by either autocorrelation or covariance methods [5].

Linear Predictive Cepstral Coefficients (LPCC) represents LPC coefficients in the cepstral domain [12]. This feature reflects the difference of the biological structure of human vocal track [9]. LPCC derives from LPC recursively as [11]

$$\begin{cases} LPCC_1 = LPC_1 \\ LPCC_i = LPC_i + \sum_{k=1}^{i-1} \frac{k}{i} LPCC_{i-k} LPC_k \,, 1 < i \leq M \end{cases} \tag{5}$$

where M is LPCC coefficients order, $i = 2, \cdots, M$.

B. Mel Frequency Cepstral Coefficients

Mel frequency cepstral coefficients (MFCC) are coefficients that describe the mel frequency cepstrum [13], [14]. In sound processing, the mel frequency cepstrum is a representation of the short-time power spectrum of a signal based on a linear cosine transform of a log spectrum on a non-linear mel scale of frequency. The mel frequency cepstrum is obtained through the following steps. The short-time Fourier transform of the signal is taken to obtain the quasi-stationary short-time power spectrum $F(f) = F\{f(t)\}$. The frequency portion of the spectrum is mapped to the mel scale perceptual filter bank using 18 triangle band pass filters equally spaced on the mel range of frequency $F(m)$. These triangle band pass filters smooth the magnitude spectrum such that the harmonics are flattened in order to obtain the envelope of the spectrum with harmonics. The log of this filtered spectrum is taken and then the Fourier transform of the log spectrum squared results in the power cepstrum of the signal.

$$Mel(f) = 2595\, log_{10}(1 + \frac{f}{700}) \tag{7}$$

At this point, the discrete cosine transform (DCT) of the power cepstrum is taken to obtain the MFCC, a tool commonly used to measure audio signal similarity. The DCT coefficients are retained as they represent the power amplitudes of the mel frequency cepstrum.

C. Bark Frequency Cepstral Coefficients

Similar to the MFCC, the BFCC warps the power cepstrum such that it matches human perception of loudness. The methodology of obtaining the BFCC is similar to that of the MFCC except for two differences [12]. The frequencies are converted to bark scale with the formula below:

$$Bark(f) = 13 \arctan(0.00076f) + 3.5 \arctan((\frac{f}{7500})^2) \tag{10}$$

where b denotes bark frequency and f is frequency in hertz. The mapped bark frequency is passed through 18 triangle band pass filters. The center frequencies of these triangular band pass filters correspond to the first 18 of the 24 critical frequency bands of hearing.

The BFCC is obtained by taking the DCT of the bark frequency cepstrum and the 10 DCT coefficients describe the amplitudes of the cepstrum.

IV. Classification

A. Nearest Neighborhood

Earlier, signal detection of cries was used to find the waveform boundaries of cries to be processed. Furthermore, these cry signals were further processed in smaller frames every 16 ms in time (or 128 samples) to accurate obtain feature parameters that could describe the frame in detail in a fashion such that it could be compared with parameters from different signals without losing the integrity of the comparison.

As the features of the infant cry signals were deconstructed into an array of LPCC, MFCC or BFCC. To generate static codebooks, codebooks whose size is independent of the signal duration, each cry signal was subdivided into 10 blocks. Within these 10 blocks, frame by frame analysis inside the blocks was performed by implementing the feature extraction algorithms over 50% overlapping frames of 16 ms that were windowed by a hamming window. The codebook for three different cry signals (1=hungry, 2=diaper and 3=attention)

$$\mathbf{C} = \begin{bmatrix} c_{11} & \cdots & c_{1N} \\ c_{21} & \cdots & c_{2N} \\ c_{31} & \cdots & c_{3N} \end{bmatrix}$$

where N=10.

The method of choosing the best codebook matches was determined by developing a cost function, whose parameter was mean square error, where the lowest cost function values were designated as the best fit [12].

The test feature is $\boldsymbol{t} = [t_1 \quad \cdots \quad t_N]$, where N=16, and classification result is

$$\text{argmin} \sum_{j=1}^{N} (t_j - c_{ij})^2 \tag{11}$$

B. Artificial Neural Network

ANN imitates how human brain neurons work to perform certain task, and it can be considered as a parallel processing network system with a large number of connections [15]. ANN can learn the rule from examples and generalize relationships between inputs and outputs, or in other words, find patterns of data. The Learning Vector Quantization (LVQ) model can implement the classification of multi-classes issue. The objective of using LVQ ANN model for baby cry causes recognition is to develop 3 feature patterns which represent the cluster centroids of each baby cry cause-draw attention cry, wet diaper cry, and hungry cry, respectively.

As we assume that different causes baby cry have different feature patterns, the objective of classification is to find out a general feature pattern which is a kind of MFCC "codebook" from example training feature data for a specific baby cry cause, such as draw attention cry, need to change wet diaper cay, or hungry cry, and etc. Then recognize the unknown cause baby cry by finding out the shortest distance between the input unknown cry word MFCC-10 feature vector and every class "codebook" respectively.

In this paper, LVQ algorithm is used to complete the baby cay caused classification, and three main baby cry causes are took into consideration, that are draw attention , diaper change needed, and hungry. Thus the LVQ neural network has 3 output classes which are corresponding to the 3 main baby cry causes: 1: Draw attention cry, 2: Diaper change needed cry and 3: Hungry cry.

For each cry word, 10-order coefficients feature is extracted. The input vector of LVQ has 10 elements as

$$\vec{\mathbf{x}} = [x_1 \; x_2 \; \cdots \; x_{10}]^{\mathrm{T}} \tag{12}$$

The weights matrix of LVQ neural network can be expressed as

$$\mathbf{W} = [\vec{\mathbf{w}}_1 \quad \vec{\mathbf{w}}_2 \quad \vec{\mathbf{w}}_3 \quad \vec{\mathbf{w}}_4] = \begin{bmatrix} w_{1,1} & \cdots & w_{4,1} \\ \vdots & \ddots & \vdots \\ w_{1,10} & \cdots & w_{4,10} \end{bmatrix} \tag{13}$$

where $\vec{\mathbf{w}}_i = [w_{\mathrm{i},1} \; w_{\mathrm{i},2} \; \ldots \; w_{\mathrm{i},10}]^{\mathrm{T}}$ represents the pattern "codebook" of Class i, the subscript of each weight coefficient $w_{\mathrm{i,j}}$ is correspond to the path from jth input layer element to ith objective output layer element.

LVQ neural network model needs to be trained using some known cause cry words to obtain the reference "codebook". The train procedure is completed by the follows steps:

1. Initialize all weight vectors $\vec{\mathbf{w}}_i(1)$. Initialize the adaptive learning step size $\mu(k) = \frac{\mu(0)}{k}, \mu(0) = 0.1$. The training iteration index starts from $k = 1$.

2.M is the number of training input feature vectors used in each iteration, for each training input vector $\vec{\mathbf{x}}(m)$, $m = 1,2,\cdots,M$, perform step 3 and step 4:

3. Obtain the weight vector subscript q such that the Euclidean distance $\left\|\vec{\mathbf{x}}(m) - \vec{\mathbf{w}}_q(m)\right\|^2$ is minimal. Save the subscript of the optimal $\vec{\mathbf{w}}_q(m)$ as $C_{\vec{\mathbf{w}}_q(m)} = q$, and $C_{\vec{\mathbf{w}}_q(m)}$ is the training output class number for this train input feature $\vec{\mathbf{x}}(m)$.

4. Update the appropriate weight vector $\vec{\mathbf{w}}_{\mathrm{q}}(\mathrm{m})$ as follows:

$$\begin{aligned} &\vec{\boldsymbol{w}}_q(m+1) = \vec{\boldsymbol{w}}_q(m) + \mu(k)\left[\vec{\mathbf{x}}(m), -\vec{\mathbf{w}}_q(m)\right], C_{\vec{\mathbf{w}}_q(m)} = C_{\vec{\mathbf{x}}(m)} \\ &\vec{\boldsymbol{w}}_q(m+1) = \vec{\boldsymbol{w}}_q(m) - \mu(k)\left[\vec{\mathbf{x}}(m), -\vec{\mathbf{w}}_q(m)\right], C_{\vec{\mathbf{w}}_q(m)} \neq C_{\vec{\mathbf{x}}(m)} \end{aligned} \tag{14}$$

where $C_{\vec{\mathbf{x}}(m)}$ is the known class number of input $\vec{\mathbf{x}}(m)$.

5. $k = k + 1$, update training step size $\mu(k) = \frac{\mu(k-1)}{k}$ for next iteration, and repeat Step2, Step3, Step 4, until $k = K$, where K is the total iteration number.

After training, $\vec{\mathbf{w}}_1$, $\vec{\mathbf{w}}_2$, $\vec{\mathbf{w}}_3$, and $\vec{\mathbf{w}}_4$ are the reference "codebook" for corresponding cause class, respectively.

V. Simulation Results

All the baby cry data files used in this paper are recorded by author in neonatal intensive care unit (NICU) of a hospital in China. The probable cry causes of each recording file are given buy experienced neonatal nurses. These 30 recording includes 8 draw attention cry wave files, 6 diaper change needed cry wave files, and 16 hungry cry wave files. Recording sampling frequency is 44.1k Hz.

Each wave file is processed to get several "cry words". One example of baby cry word detection result is shown in Fig. 1, 5 "hungry cry words" have been detected. One can

find that the short duration and low power signal segment care removed from the cry signal detection.

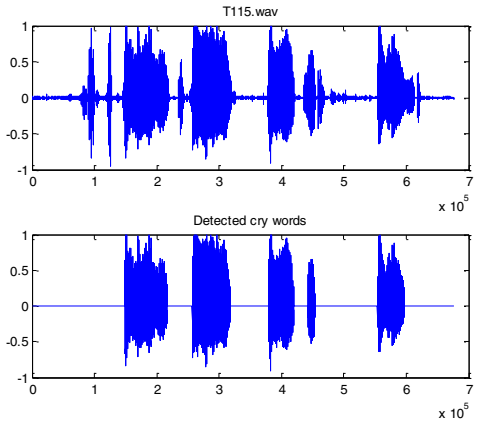

Figure 1. Baby cry signal and detected cry word.

Figure 2 shows "Draw attention cry words" MFCC-10 feature of 4 different babies. The features from one infant are similar to the other infants for the same cause or reason.

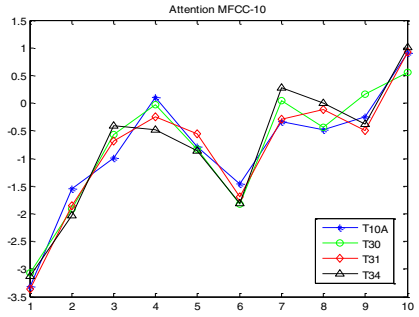

Figure 2. "Draw attention cry words" MFCC-10 features of 4 different babies.

Figure 3 shows "Diaper change needed cry words" MFCC-10 features of 4 different babies. The results show certain degree of similarity for the same reason 'diaper' from different infants.

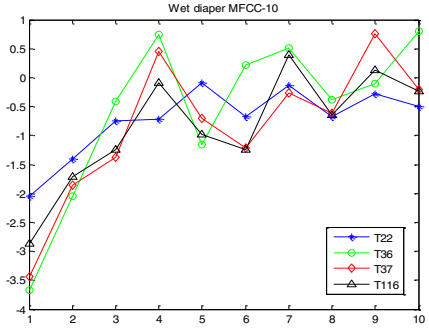

Figure 3. "Diaper change needed cry words" MFCC-10 features of 4 different babies.

Figure 4 shows "Hungry cry words" MFCC-10 features of 4 different babies. We can find that the MFCC obtained from 'hungry' is quite different from 'attention' cry and 'attention' cry signals.

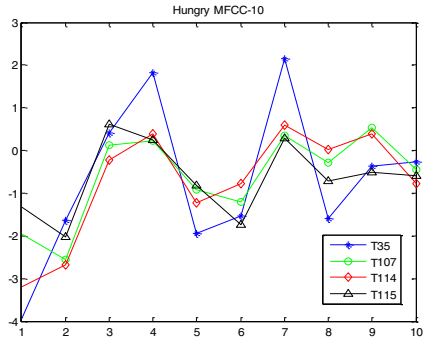

Figure 4. "Hungry cry words" MFCC-10 features of 4 different babies.

The classification performance can be defined as:

$$\text{correct } rate = \frac{m}{M} \times 100\% \qquad (15)$$

where m is the number of right classification, and M is the total number of test cry words.

TABLE I. INFANT CRY RECOGNITION CORRECT RATE BY USING DIFFERENT FEATURES

Features	LPC	LPCC	MFCC	BFCC
Nearest neighborhood	63.84	47.95	63.89	65.22
ANN LVQ	54.55	51.88	60.45	76.47

Highest recognition correct rate for infants cry application is achieved at 76.47% by using BFCC-10 feature.

VI. CONCLUSION

This paper presents infant cry detection method, then the different features are extracted from the cry signals. Nearest neighborhood and neural network methods are used to recognize the infant cry signals. The simulation results show that classification rate is around 70%. It is believed that there are patterns that distinguish the meaning of infant cries. Future work can include determining the meanings behind infant cries based on not only audio cues but also upon observation and deductive logic as well.

REFERENCES

[1] Lederman, D. "Automatic Classification of Infants' Cry". M.Sc dissertation, Ben-Gurion University, Beer-Sheva, Israel, 2002.

[2] K. Kuo, "Feature Extraction and Recognition of Infant Cries," 2010 IEEE Int. Conf. on Electro/Information Technology, Normal, Illinois.

[3] *Cry Translator*. Biloop Technologic,S.L., 2009. Web. 05 Mar. 2011. <http://www.crytranslator.com>.

[4] Lichuan Liu, Kevin Kuo and Sen M. Kuo, "Infant Cry Classification Integrated ANC System for Infant Incubators", appear on IEEE International Conference on Networking, Sensing and Control, April 10-12, Paris, France, PP. 383-387.

[5] L. Liu, Kevin Kuo, "Active Noise Control Systems Integrated with Infant Cry Detection and Classification for Infant Incubators", Acoustic 2012, HongKong, pp.1-6

[6] Várallyay Jr., György, "Future Prospects of the Application of the Infant Cry in the Medicine," Periodica Polytechnica Ser. El. Eng. vol. 50, no. 1-2, (2006), pp. 47-62.

[7] G. Buonocore, and C.V. Bellieni, Neonatal Pain, Suffering, Pain and Risk of Brain Damage in the Fetus and Newborn. Springer, 2008

[8] Kondoz, A. M., Digital Speech, John Wiley & Sons Ltd, West Sussex, England, 2004.

[9] Hayes, Monson H., Statistical Digital Signal Processing and Modeling, John Wiley & Sons Ltd, West Sussex, England, 1996.

[10] Owens, F. J., Signal Processing of Speech, Macmillan Press LTD, London, England, 1993.

[11] Rowden, Chris, Speech Processing, McGrawHill Book Company, London, England, 1992.

[12] Gold, B and Morgan, N. Speech and Audio Signal Processing. New York: John Wiley & Sons, 2000.

[13] Md. Sahidullah, G. K. Saha, Design analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition, Speech Communication, Volume 54 Issue 4, May, 2012.

[14] B. Martin, Extraction of feature from the acoustic activity of RPW using MFCC, Recent Advances in Space Technology Services and Chimate Change (RSTSCC), 2010.

[15] F. M. Ham, I. Kostanic, Principles of Neurocomputing for Science and Engineering, McGraw-Hill, INC. 2001.