AN ASSESSMENT OF THE PERCEPTUAL ROLE OF INDIVIDUAL ACOUSTIC (INFANT CRY)

Muhammad Hafiz Abdullah @ Gerugu Randi (2004617147)

Faculty of Electrical Engineering, MARA University of Technology (UiTM), Shah Alam, Selangor, Malaysia

Abstract – This paper presents the build up of feature extraction for infant cry, with the objective to find new algorithm for three types of infant with three types of cry: normal and operate infant born with less risk to having disease and also normal infant born with potential to having serious illness. Linear prediction cepstral coefficient technique is used to obtain acoustic characteristics. The results show that we can extract the information from infants cry signal to produce a characteristic for feature extraction.

1 Introduction

Human in the age from he or she birth until can speak (12 months or more) was call as infant. Infant is condition where children unable to speak. So, basically infant will cry in order to express a variety of feelings including hunger, discomfort, over stimulation, boredom, wanting something, or loneliness with his or her parent. The infant can make different sound of cry depending on his or her physical and psychological state. Based on human and animal studies, it is known that the cry is related to the neuropsychological status of the infant. The studies had led to the development of conceptual models that describe the anatomical and physiologic basis of the production and neurological control of the cry [1]. According to the specialists, babies' crying wave carries useful information, as to the physical state of the baby, as well as to detect possible the physical pathologies, mainly cerebral, from very early stages [1], [2], [3].

The purpose of this study was to assess the role of various acoustic parameters on the perception of infant cries by adults using computer generated cries. Based on a number of real cries, linear prediction cepstral coefficients analysis was used to create a set of parameters for each cry, which were then altered in a controlled way to synthesize artificial cries. The correlation of the individual acoustic features with perceived qualities provides a reliable assessment of their perceptual role.

The current state of research on why and how the infants cry, what kind of information might be contained in the cry, and how the cry affects and is perceived by adults was presented, all in relation to the cry acoustics, which are of major interest for this study. For a valid assessment of the value of any acoustic parameter in a broader context, it is critical to understand all the stages involved, and how each one might influence the acoustics.

The most well documented acoustic feature with respect to its perceptual effects in adults is the mean fundamental frequency of the cry. The most unlooked at, yet very promising because of other research in related fields, is jitter (perhaps perturbations of F_0 in general). Another very interesting feature for which very few data exist is the long-term variation of the fundamental frequency (melodic contour), and in particular the time it takes for the F_0 to reach its maximum, known as rise time.

2 Infant Cry

When an infant crying, his or her lungs work like a power supply of the cry production system. The glottis supplies the input with the certain pitch frequency (F_0). The vocal tract, which consists of the pharynx and the mouth and nose cavities, works like a musical instrument to produce a sound. In fact, different vocal tract shape would generate a different sound. To form different vocal tract shape, the mouth cavity plays the major role. To produce nasal sounds, nasal cavity is often included in the vocal tract.

The nasal cavity is connected in parallel with the mouth cavity. The glottal pulse generated by the glottis is used to produce vowels or cry sounds. And the noise-like signal is used to produce consonants or unvoiced sounds.

A child's pitch frequency can go as high as 400 Hz. This glottal pulse excites a vocal tract cavity and produces a vowel (or voiced) sound

3 The Feature Extraction Procedure

The goal of the feature extraction procedure is to describe the audio cry signal in the features space by means of a sequence of observation vector (O).

Various features such as LPC-based features (LPC, LPCC, etc.), MFCC and filters bank

coefficients may be used [4]. The steps required for the feature extraction procedure are described following. These steps are somewhat general in the sense that not all of them are necessary to perform.

For 9 samples with 3 reason cries (hunger, pain and pleasure) of 3 different infant was successfully analysis in the automatic recognition system. First infant was normal infant born and less potential having serious disease in the future with 3 condition reason of cry (hunger, pain and pleasure). We name first infant as Infant A. The second baby (namely as Infant B) was operation infant born and also less potential to having serious illness with 3 similar reason of cry. The last infant (call as Infant C) was normal infant born but having potential to get serious disease in the future with 3 same reason of cry [5] [6].

The first step of the feature extraction procedure is filtering the signal with a pre-emphasis filter with a transfer frequency:

$$P(z) = 1 - 0.97z^{-1} \qquad (1)$$

The main purpose of this step is to effectively eliminate the spectral contributions of the larynx and lips so that the analysis can be asserted to be seeking parameters corresponding to the vocal tract only.

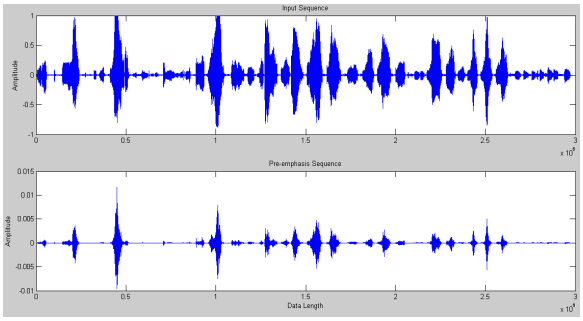

Figure 1 - Infant A in pleasure condition

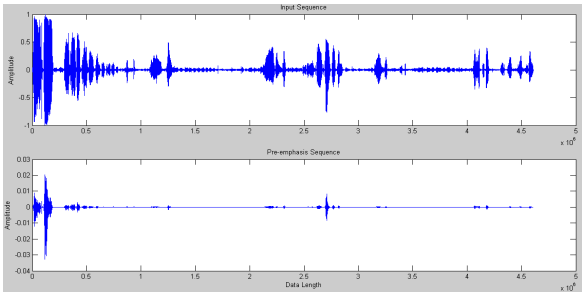

Figure 2 – Infant B in pleasure condition

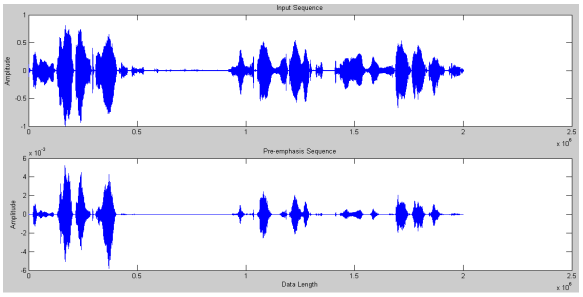

Figure 3 – Infant C in pleasure condition

From figure 1 until figure 3, the figure divides into two where input sequence is mean full wave cry signal before the filtering. The peak and size of the wave amplitude is high and large because the noise contribution. For this reason, the pre-emphasis or filtering are necessary to eliminate the noise and leave only important data information wave form. The size of data length becomes small after the filtering and also the peak of amplitude given by small value. The voice and unvoiced part can be differentiating directly from this figure. From figure, unvoiced part shown as part that only has straight line without any wave form. Before filtering, it hard to say the straight line is unvoiced because we can see waveform in it where it contribute by noise. After the pre-emphasis process, only the important information data without noise will be analysis and become input for linear prediction cepstral coefficients.

The cry signals are non-stationary. Hence short term analysis must be used in which the signal is divided into quasi stationary overlapping frames and each frame is multiplied by a suitable Hamming window. The frame length is pre-chosen in order to ensure quasi-stationary of the framed signal on the one hand and on the other hand, to ensure that each frame will include at least one period of the fundamental frequency and to decrease the amount of computations as much as possible. In cry signals the fundamental frequency is known to have a wide range of possible values (about 200- 2500 Hz). Consequently, the appropriate practical range for frames duration should be around 5-25 millisecond.

In this step the first p +1(p – LPC order) autocorrelation coefficients are estimated out of each windowed frame of the signal.

Using the autocorrelation coefficients, a vector of LPC coefficients is calculated for each windowed frame with the Levinson-Durbin recursive algorithm [7].

5 Results and Discussion

There are several ways to estimate the pitch of a voiced sound: autocorrelation method (refer to “4 the feature extraction procedure” for explanation

and understanding), average magnitude difference function and cepstrum.

The autocorrelation of a stationary sequence $x(n)$ is defined as

$$R_x(\tau) = \langle x(n)x(n+\tau)\rangle = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x(n+\tau)$$

where τ is termed the lag. Auto means self or from one signal, and correlation means relation between two samples. An autocorrelation is the average correlation between two samples from one signal that are separated by τ samples. It should be noted that the upper limit in the summation will be less than N−1 when τ is positive, and the lower limit will be greater than 0 when τ is negative. Thus, the autocorrelation can be rewritten as

$$R_x(\tau) = \frac{1}{N}\sum_{n=0}^{N-1-|\tau|} x(n)x(n+|\tau|) \qquad (3)$$

Because the number of items in the summation decreases as τ increases, the envelope of the autocorrelation decreases linearly as τ increases. In some cases, to prevent this tapering, autocorrelation is defined alternatively as

$$R_x(\tau) = \frac{1}{N-|\tau|}\sum_{n=0}^{N-1-|\tau|} x(n)x(n+|\tau|) \qquad (4)$$

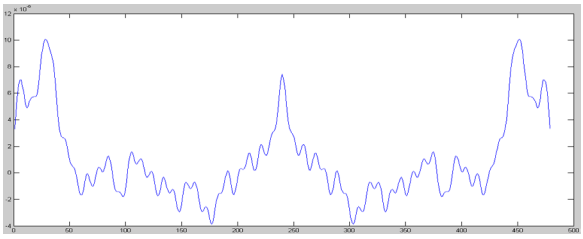

Figure 4 – Infant A in pleasure condition

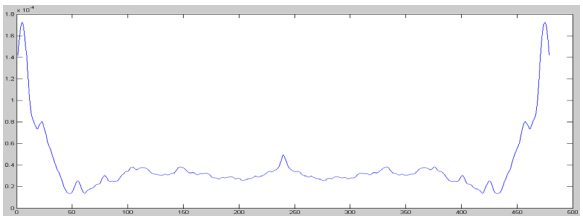

Figure 5 – Infant B in pleasure condition

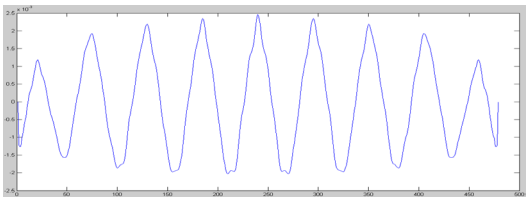

Figure 6 – Infant C in pleasure condition

The average magnitude difference function uses the following property. Suppose that a signal $x(n)$ is periodic with period T. Then the difference between two samples

$$Diff(k) = x(n) - x(n+k) \qquad (5)$$

will be zero for $k = 0, \pm T, \pm 2T$ and so on. Because a voiced sound is not exactly periodic, the short time average magnitude difference function (AMDF) is defined as

$$AMDF(k) = \frac{1}{N-k}\sum_{n=0}^{N-1-k} x(n)x(n+k) \qquad (6)$$

for positive k.

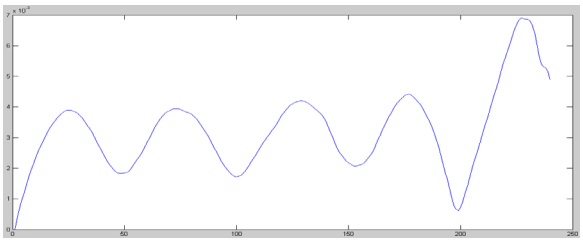

Figure 7 – Infant A in pain condition

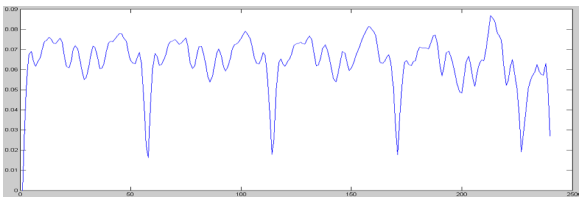

Figure 8 – Infant B in pain condition

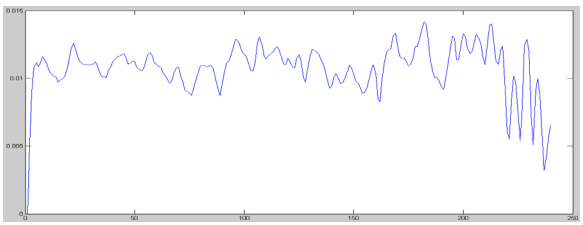

Figure 9 – Infant C in pain condition

Average magnitude difference functions divide the size of frame from autocorrelation analysis into half (equal to 250 Hz). All 9 figures (figure 7 until figure 9 are example) as expected, shown different pattern of wave form. But signal wave form for average magnitude difference functions having similar form of wave signal for autocorrelation analysis. Only value of amplitude and the range of maximum peak to minimum peak are changing in size. For average magnitude difference functions also use comparison in amplitude peak to peak and it range to differentiate the type of infant born and theirs condition. This is the second characteristic for feature extraction that can be use.

There are two kinds of cepstrum: the real cepstrum and the complex cepstrum. Only the real cepstrum is explained here. Suppose that $x(n)$ is an infant cry signal. The magnitude spectrum $|X(k)|$ is obtained by computing the magnitude of the DFT of $x(n)$. The real cepstrum is defined as the inverse discrete Fourier transform of the logarithm of the magnitude response, for example,

$$c(n) = IDFT\{\log|X(k)|\} \tag{7}$$

where $c(n)$ is termed the cepstrum and log is the natural logarithm.

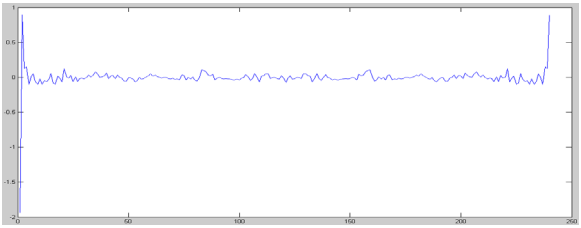

Figure 10 – Infant A in hungry condition

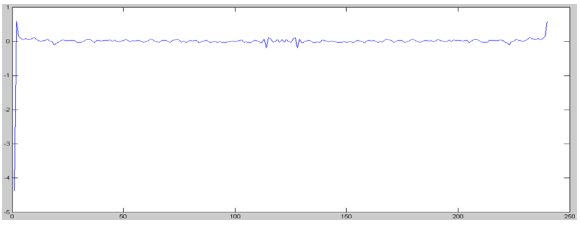

Figure 11 – Infant B in hungry condition

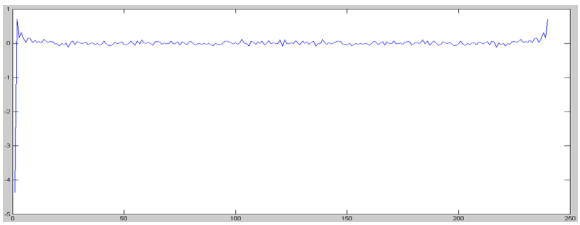

Figure 12 – Infant C in hungry condition

In cepstrum analysis, both amplitude and frequency were set into fix and same value for all type of infant and condition. As we seen, it looks similar pattern wave form for all patterns. In detail, it having a different value of start and ending point.

Voiced/unvoiced detection also can be use to find characteristic of feature extraction besides using pitch detection. One important task is segmentation and labelling of each segment as voiced of unvoiced. To identify whether the cry segment is voiced or unvoiced cry, spectral flatness measure, energy, and zero crossing rates are most widely used. The spectral flatness makes use of the property that the spectrum of pure noise is expected to be flat. In other words, the spectrum of unvoiced section is flat and the spectrum of voiced section is less flat. The spectral flatness measure (SFM) is given by

$$SFM = \frac{G_m}{A_m} \tag{7.8}$$

G_m is the geometric mean of the magnitude spectrum and is determined by multiplying all the spectral lines together and raising the final product to one over the total number of spectral lines. A_m is the arithmetic mean of the magnitude spectrum and is obtained by taking the sum of the spectral lines divided by the number of spectral lines.

$$SFM = \frac{\left(\prod_{k=0}^{N-1} X_j(k)\right)^{\frac{1}{N}}}{\frac{1}{N}\sum_{n=9}^{N-1} X_j(k)} \tag{8}$$

$X_j(k)$ is the magnitude of the N-point DFT of the j^{th} frame of the speech signal. The spectral flatness measure ranges from 0.9 for a white noise to 0.1 for a voiced signal. The threshold is usually chosen to be 0.35 ~ 0.48.

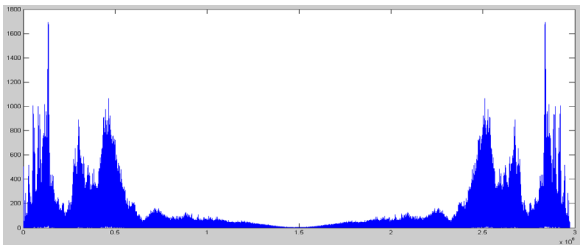

Figure 13 – Infant A in pleasure condition

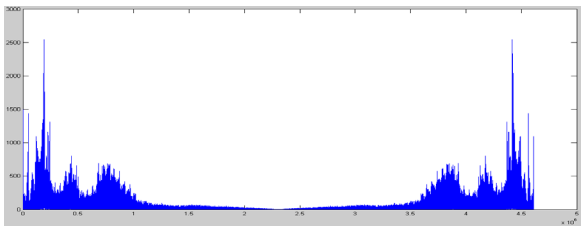

Figure 14 – Infant B in pleasure condition

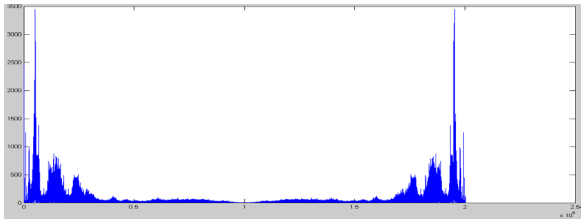

Figure 15 – Infant C in pleasure condition

Voiced/unvoiced detection for all type of infant and condition are shown from figure 13 until figure 15. Certain range frequency having flat wave form where that wave form is unvoiced form. For "infant C" in pleasure condition (figure 15), the unvoiced segment start from 0.5 MHz until 1.5 MHz. This unvoiced part can be vanish or set it into flat waveform. Different type of infant and condition will produce different point of frequency for unvoiced part. Use this different value of frequency to become it one of most characteristic for feature extraction.

6 Conclusions and Future Work

Each time infant meaning he or she will showing his or her condition or need. There many method to extract the information from cry signal. Linear prediction cepstral coefficients is one of the method that widely use. For the future work, this system can add more function and extract more information from infant cry.

The more feature extraction characteristic that been use will make the system work efficiently. For this paperwork, it only had shown four methods that easy to understanding and use.

In future work, more method for the comparison that can be use likes auditory analysis, time domain analysis, frequency-domain analysis and computer-based analysis. Also from the waveform of cry, more information can be extract like gender, age, weight and type of disease.

REFERENCES

[1] Jose Orozco, Carlos A. Reyes-Garcia, *Implementation and Analysis of Training algorithms for Classification of Infant Cry with Feed-forward Neural Networks,* Instituto Nacional de Astrofisica Optica y Electronica (INAOE), Lius Enrique Erro # 1, Tonantzintla, Puebla, Mexico, 4-6 September 2003, pp 271 – 276

[2] Jose Orozco, Carlos A. Reyes-Garcia, *Mel-Frequency Cepstrum Coefficients Extraction from Infant Cry for Classification Of Normal and Pathological Cry with Feed-forward Neural Networks,* Instituto Nacional de Astrofisica Optica y Electronica (INAOE), Lius Enrique Erro # 1, Tonantzintla, Puebla, Mexico, 2003, pp 3140 - 3145

[3] Orion F. Reyes-Galviz*, Carlos Alberto Reyes-Garcia**, *a System for the Processing of Infant Cry to Recognize Pathologies in Recently Born Babies with Neural Networks,* *Instituto Tecnologico de Apizaco, Av. Tecnologico S/N, Apizaco, Tlaxcala, 90400, Mexico, **Instituto Nacional de Astrofisica Optica y Electronica, Lius E. Erro 1, Tonantzintla, Puebla, 72840, Mexico, 20-22 September 2004

[4] Munehiro Namba, Hiroyuki Kamata, Yoshihisa Ishida, *Neural Networks Learning with L1 Criteria and Its Efficiency in Linear Prediction of Speech Signals,* School of Science and Technology, Department of Electronics and Communications, Meiji University, Japan, pp 1245 – 1248

[5] Cohen, A. and Zmora, E. (1984), *Automatic classification of infants' hunger and pain cry,* In V. Cappellini and A. G. Constantinides (Eds.), Digital Signal Processing | 84, pages 667{672, Elsevier Science Publications, Amsterdam.

[6] Colton, R. H., Steinschneider, A., Black, L., and Gleason, J. (1985), *The newborn infant cry: Its potential implications for development and SIDS,* in B. M. Lester and C. F. Z. Boukydis (Eds.), Infant Crying, chapter 6, pages 119{137. Plenum Press, New York.

[7] Atal, B. S. and Hanauer, S. L. (1971), *Speech analysis and synthesis by linear prediction of the speech wave,* The Journal of the Acoustical Society of America, 50(2):637{655.