Volume 6, Number 2 December 2008 ISSN 1881-4581

Journal of
eHealth Technology and Application

Supported by ITU-D SG-2 Q14

Published by
Tokai University
National Institute of Information and Communications Technology

Journal of eHealth Technology and Application

Supported by ITU-D SG-2 Q14

Article adoption conditions are as follows;

1. All scientific papers (original articles) are peer reviewed by above-listed specialists.
2. Brief communicationa are selected by the Editors-in-Chief as a technical paper.
3. From the standpoint of the international policy, the Rapporteur of the Q14(ITU-D SG2 telecommunications for eHealth) can request the author to contribute her/his article to the report of ITU-D. In that case, another Copyright Assignment Agreement for the ITU publication has to be assigned separately from this Journal.

Journal of eHealth Technology and Application
Volume 6, Number 2
December 2008
ISSN : 1881-4581

EDITOR: ISAO NAKAJIMA, MD, PhD

Professor of Emergency and Critical Care
Director of Telemedicine & eHealth
Tokai University School of Medicine
Shimokasuya 143, Isehara-shi
Kanagawa, 259-1143 Japan
Tel: +81-463-91-3130
Fax: +81-463-91-0780
E-mail: jh1rnz@aol.com

Preface

Asif Zafar Malik, MD, FCPS, FRCS
Telemedicine & e- Health Training Center
Holy Family Hospital, Rawalpindi, Pakistan

The number of natural disasters around the world is on the rise, killing by the thousands, affecting lives of millions of people, and causing billions of dollars in economic damage. Most deaths occur in poor countries, whereas the largest amount of economic damage happens in rich countries. In 2004, the Indian Ocean tsunami claimed 226,000 – 250,000 lives, while in Pakistan there were 73000 deaths in 2005. In Katrina there were 1300 dead with over $ 125 billion losses. Additional threats such as global warming, environmental degradation and rapid urbanization make millions of people more vulnerable to natural hazards especially for those living in remote and isolated areas.

When disaster strikes communications links are often disrupted, but for disaster relief workers these links are essential to answer critical questions such as how many people have been injured or died, where they are located and the medical help needed. Disaster response to mass-casualty incidents represents one of the greatest challenges to a community's emergency response system. One consistent challenge for disaster response is communication and information management.

Natural and man-made hazards cannot be entirely prevented, but ICTs can help reduce their impact and limit damage. Telecommunications are critical at all stages in prevention, preparation, response and relief efforts. Telecommunications are critical in the immediate aftermath of a disaster, ensuring timely communications and the flow of information needed by governments and relief agencies to organize rescue operations and provide medical assistance. Reconstruction of disrupted telecommunication networks is also vital.

Telemedicine applications have been successfully demonstrated during disaster situations. Telemedicine was first applied in disasters during the mid-1980s The National Aeronautics and Space Administration (NASA) first used telecommunication technology to furnish disaster aid following the devastating 1985 earthquake in Mexico City. The U.S.–U.S.S.R. Space Bridge project was employed after the Armenian earthquake in 1988. The U.S. armed forces have provided mobile health and telemedicine services during Hurricane Hugo, Haiti and Bosnia in 1990's. Simple systems for disaster telemedicine can often deliver. Satellites offer a method of long-distance communication when other means, such as land lines or cellular telephone services, are destroyed by disaster as seen in Pakistan earthquake in 2005. However, the advantages of satellites come at a great cost.

International Telecommunications Union (ITU) makes invaluable contribution to disaster management by deploying temporary telecommunications / ICT solutions to assist countries affected by disasters. ITU provided to Government of Pakistan 40 Inmarsat Satellite modems during the earth quake of October 2005. 15 modems were provided to Telemedicine & E-health training center Holy Family hospital, Rawalpindi. Mobile Telemedicine units were set up in NWFP Province and Azad Kashmir. These were stationed at Shohal Najaf, field hospital Balakot and Hattian Bala, to cater the emergency and diagnostic medical needs of the affected of the earthquake. Mobile telemedicine centers in step down hospitals provided follow-up of all patients shifted to in a remote hospitals. This clearly demonstrated utility in the follow up of trauma patients remotely and assessment of missed injuries. This experience of complementing Emergency relief work with mobile Telemedicine units is extremely valuable and can easily be replicated and deployed on urgent basis in wake of disasters.

Telecommunications can save lives in disaster situations; regulatory barriers can make it difficult to use the necessary equipment. ITU was a driving force in drafting and promoting the Tampere Convention. It allows relief workers to make full use of life-saving communication tools. The Tampere Convention calls on States to waive regulatory barriers that impede the use of telecommunications. These barriers include licensing requirements to use frequencies, restrictions on importing equipment and limits on the movement of humanitarian teams.

The special issue on Telecommunications for disaster and Emergency Medicine incorporates research articles in this field. This experience will be invaluable in preparing to deal with disasters.

Journal of eHealth Technology and Application

Volume 6 Number 2 December 2008

I: Telecommunication for Disaster and Emergency Medicine

II: Original Articles

A Low Bit Rate Speech Coder using Segmental Sinusoidal Model for Disaster and Emergency Telemedicine

Florentinus B. Setiawan[1,2], Soegijardjo Soegijoko[1], Sugihartono[1], Suhartono Tjondronegoro[1]

[1]Institut Teknologi Bandung, Indonesia, [2]Soegijapranata Catholic University, Indonesia
email: fbudisetiawan@yahoo.com, f_budi_s@unika.ac.id

***Abstract*— In general, a communication system during and after disaster needs to work properly at a relative very low data rate. This performance is also required in emergency telemedicine, because of limited channel capacity. Limited infrastructure available during and after disasters, and emergency conditions reduce the communication system into its minimum capacity. To establish a communication connection during and after disaster, it needs a speech coder that can function properly at low bit rate. The proposed speech coder is a low complexity coder that should be able to function properly. Therefore, the large number of communication connection can be handled using the limited transmission channel. The low complexity and low bit rate speech coder can be realized using segmental sinusoidal model, so that, the speech signal can be represented as a combination of sinusoidal signal with infinite combination of amplitude, frequency and phase. The segmental sinusoidal model extracted from the periods and the peaks of speech signal along one frame. This model works based on peak-to-peak quantization that detects the positive peaks and the negative peaks. Thus, the time distance and magnitude difference between the consecutive peaks can be easily extracted. In this paper, we describe the proposed method called segmental sinusoidal model to encode a speech signal. A low bit rate can be obtained by sending the information of periods and peaks. This coder is also combined with the waveform interpolation and the use of look-up tables. The resulted maximum mean opinion score (MOS) of the synthesized speech signal is 3.8. With this MOS test score, the human perception due to the synthesized signal is fairly good. The bit rate of the coded signal is 4 kbps at less than 10 MIPS complexity. It is therefore expected that the proposed segmental sinusoidal model and 4 kbps coder will be suitable for disaster and emergency telemedicine applications.**

***Index Terms*— analysis, disaster, frequency, interpolation, peak, period, segmental, sinusoidal, synthesis, telemedicine.**

I. Introduction

Communication is vital during a disaster and emergency telemedicine. It is important to think about what pieces of the current communication system might be inoperable during the time of a disaster. What additional communication needs might be necessary are designated a trained spokesperson and an alternate to communicate with the media and community, update employees' emergency contact information, utilize an Emergency Internet or Intranet site and utilize an Emergency Call Centre. During a disaster, an organization's ability to communicate accurate and consistent information increases the perception and ability of the business to effectively, handle the crisis. It is also very helpful in reducing the anxiety level of everyone involved.

Communication system on emergency condition during disaster and emergency telemedicine needs to work properly at very low data rate. Limited communication equipment available at disaster and emergency telemedicine reduce communication system performance into minimum capacity. To make a communication connection, it needs a speech coder that it can work in low bit rate. Thus, the large number of communication connections can be handled on the limited transmission channel.

The proposed speech coder is operated at low bit rate (4 kbps), with low complexity. Sinusoidal model based is used to develop a low bit rate speech coder. The sinusoidal model is applied under assumption that speech signal have quasi-periodic characteristics. This model is powerful for keeping perception quality, especially to hold the speech synthesis signal periodicity. Speech signal can be represented as a combination of sinusoidal signal with infinite combination of amplitude, frequency and phase. For peak-to-peak based quantization, the consecutive positive and negative peak signals are detected. Then time distance between peak to peak would be quantized. In this paper, we explain a new method to quantize the speech signal which is segmented into peak to peak based on sinusoidal modeling. The part of signal between positive peak and following negative peak or vice versa is estimated as a half period of sinusoidal signal. Magnitude between peaks is assumed twice the estimated cosine amplitude.

Speech signal can be modeled as sinusoidal signal [1][2] over the frame with length of 15 ms until 30 ms. Sinusoidal components are extracted from sinusoidal

parameters [3] and sent to receiver. The number of sinusoidal components is between 40 until 60 for synthesis signal generation. The proposed sinusoidal model is segmental sinusoidal model [4]. Signal is fetched over the variable segment. The segment length depends on maximum and minimum peaks of the speech signal. Peaks mean the maximum values or the minimum values of signal over a frame. Thus, one segment means a part of signal between a maximum peak and a consecutive minimum peak or a part of signal between a minimum peak and a consecutive maximum peak. One segment of signal between maximum peak and consecutive minimum peak can be modeled as a half period of cosine signal from $\omega=0$ until $\omega=\pi$. Then the segment of signal between minimum peak and consecutive maximum peak can be modeled as a half period of cosine signal from $\omega=\pi$ until $\omega=2\pi$. Signal components of k signal frequency with highest energy, is used to represent the signal over the frame. Reconstructed signal is equal with the original signal, if signal components k is infinite. The more k sinusoidal components, the more accurate reconstructed signal is obtained. The reconstructed signal can be written as [1][5] :

$$\tilde{s}(n)=\sum_{k=0}^{K-1} a_k \cos(\omega_k(n)+\phi_k(n)) \qquad (1)$$

$$0 < K \leq \infty$$

Where a_k represents the signal amplitude, $\omega_k(n)$ represents angle-frequency and $\phi_k(n)$ represents phase of the k-th sinusoidal signal. Based on this model, signal can be represented as k sinusoidal signal.

The paper is organized into six sections. After this introduction section, the second section explains the sinusoidal modeling of the speech signal. The next section describes encoder design, followed by the forth section of the decoder design. The fifth section explains the experimental results and the last section is conclusion of this paper.

II. Sinusoidal Model Of Speech Signal

Sinusoidal speech signal modeling can be implemented on speech signal coding, as shown in Sinusoidal Transform Coding (STC)[6][7][8]. Sinusoidal transform coding process is fetching some sinusoidal signal components which have the highest amplitude than other components. Some sinusoidal signal components appear in frequency domain as the part of the signal spectral, which have the highest magnitude. The number of the sinusoidal signals to represent the signal on a time interval called as frame are 40 until 60 sinusoidal signal components.

A. Sinusoidal signal on the Fixed Segment

The speech signal is fetched every fixed time interval with length of one frame. On the frequency domain signal representation, shown in fig.2. that there are some outstanding frequency components. The outstanding frequency components are used to represent the signal over one frame. Fig.1 shows an example of the saw-tooth signal with fundamental frequency of 66.67 Hz on 8 kHz frequency sampling. On one frame contains 240 signal samples.

The less number of harmonic signals to represent the speech signal can be realized for the shorter time segment of signal. If one segment is a block of signal between maximum peaks and consecutive minimum peaks, or minimum peaks and consecutive maximum peaks. In this condition, the length of the signal segment is varying, depend on the signal fluctuation.

By using the discrete Fourier transform, the saw-tooth signal can be represented in frequency domain.

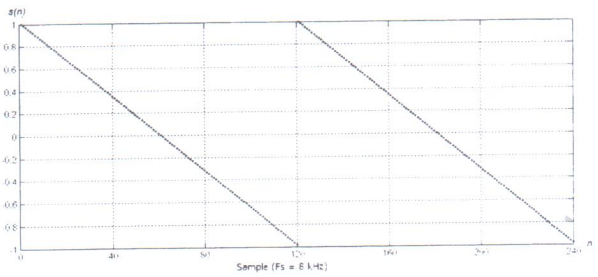


Fig. 1. Saw-tooth Signal

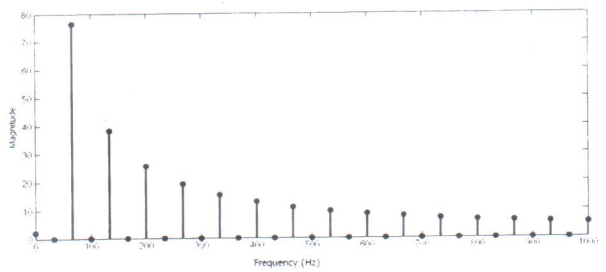


Fig. 2. Frequency domain representation of the saw-tooth signal.

B. Sinusoidal Model on the Varying Segment

The proposed model is representation of the signal into two sinusoidal signal components using varying segment. The signal fluctuation will reproduce some maximum and minimum peaks. The part of signal between the consecutive peaks contains infinite harmonic components. Human hearing perception system is sensitive with the signal periodicity level. If the signal peaks is kept, the signal periodicity will kept, then the hearing perception will increase. On the proposed method, the signal peaks position is kept and the other part of signal is reconstructed using DC-offset component and fundamental signal. Signal modeling is done by fetching signal on certain length, called as a frame. In this research, the length of frame is 30 ms or 240 signal samples on 8 kHz frequency sampling. A frame contains some

maximum and minimum peaks that change every time depends on the speech signal source.

The saw-tooth signal as shown in previous example can be assumed as a part of the triangle signal. For the first until the 120-th signal sample can be assumed as a half period of the triangle signal with period length of 240 samples. Fig.3 shows that in one signal segment contains one dominant frequency component compared with the other frequency. This component is the first harmonic signal, or the fundamental frequency of the triangle signal. By using the sinusoidal method on fetching the dominant frequency, the model is only use the first harmonic of signal to represent the speech signal. Fig.4 shows the signal reconstruction by using the first signal harmonic, as a cosine signal. For n=0,1,2,.. and k=0,1,2,3... the signal is represented as $f(n)=\cos(\pi(n-120k)/120)$.

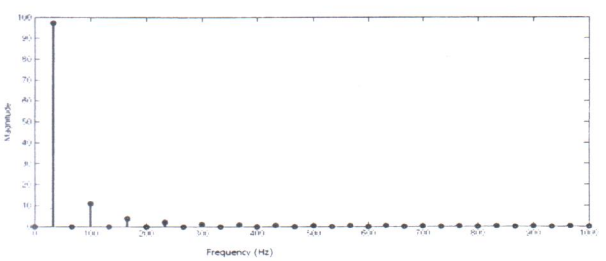


Fig. 3. The frequency domain representation of the triangle signal.

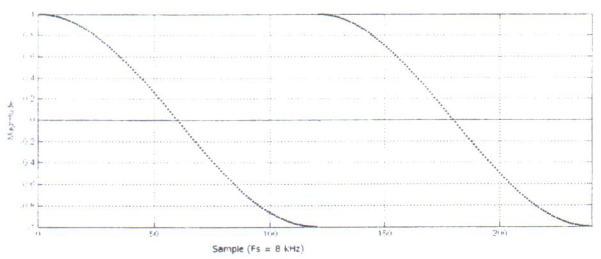


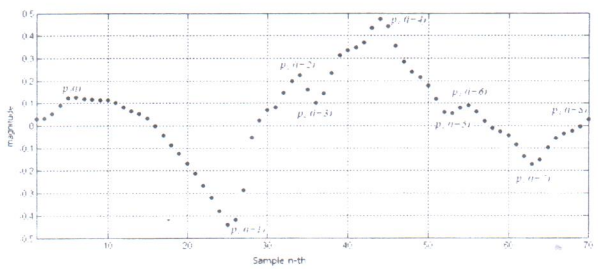
Fig. 4. The saw-tooth signal reconstruction using a half period of the triangle harmonic signal.

C. Sinusoidal Model on The Segment Between Consecutive Peaks

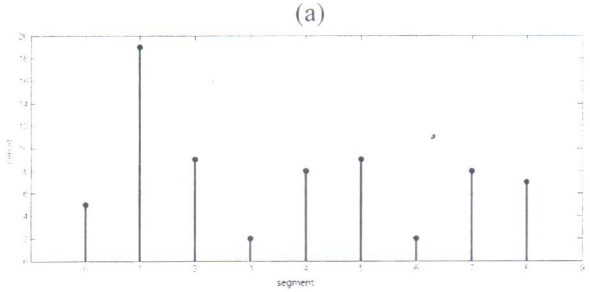
Sinusoidal model can be developed to obtain less parameters, so that the signal data rate can be reduced. This model is called as segmental sinusoidal model. By using this model, there are two harmonic signal to estimate the original signal between two consecutive peaks, (maximum to minimum or minimum to maximum). Peak means minimum peak or maximum peak on the frame. Therefore, one segment means part of signal between maximum peak and consecutive minimum peak or part of signal between minimum peak and consecutive maximum peak.

Time distance between i-th maximum peak and consecutive minimum peak called as period information, and denoted by $p_d(i)$. Maximum peak or minimum peak called as peak information, and denoted by $p_k(i)$. Peak information is obtained by detecting the maximum peaks and minimum peaks over the frame observed. Period information is obtained by counting the time distance between the consecutive peaks.

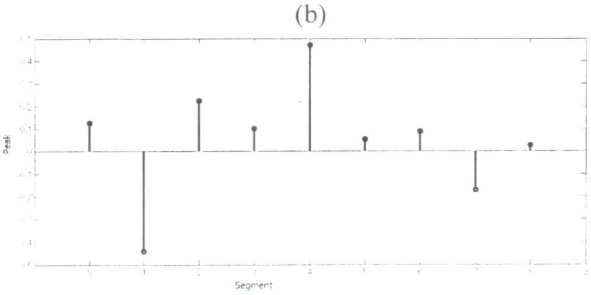
The first step to obtain the signal parameters using segmental sinusoidal model is put one frame with length of 30 ms. Next step is marking the signal peaks, both of maximum and minimum peaks. Time distance between i-th peak and $(i+1)$-th peak is same as a half period of the estimated signal, $p_d(i)$. The process is implemented for the next peaks, so that the train of period information and peaks information are obtained. Dynamic range of period information and peak information are less than the signal dynamic range. Thus, the number of signal bit to send $p_d(i)$ and $p_k(i)$ is less than the original signal.



(a)



(b)



(c)

Fig. 5. Parameter extraction by using the segmental sinusoidal model.
(a). original signal
(b). period information
(c). peak information

D. Segmental Sinusoidal Model for Signal Analysis

The proposed method is a process on time domain. On the extreme waveform coding, signal is fetched on its peaks [9]. On the one frame with length of N, there are M maximum peaks and L minimum peaks. On this frame, there are large number sinusoidal signal components. It

can be written as :

$$s(n)=\sum_{k=0}^{K} a_k \cos(\omega_k(n)n+\phi_k(n)) \quad (1)$$

for : $0 \le k \le K\text{-}1$, $n = 0,1,2 \ldots N\text{-}1$, $K \le N$

The first and second coefficients (k=0 and k=1) are used as components to reconstruct the estimated signal from one maximum peak until the consecutive minimum peak or from one minimum peak until the consecutive maximum peak. This is the equation for estimated signal :

$$s(n)=a_0+a_1\cos(\omega_1(n)+\phi_1(n)) \quad (2)$$

Part of signal from maximum peak until the consecutive minimum peak can be written as :

$$s_{pv}(n)=a_0+a_1\cos\omega_1(n) \quad (3)$$

for $n = 0,1,2 \ldots N\text{-}1$

Where a_0 is the DC-offset and a_1 is the fundamental signal.

$$a_0=\frac{p_k(i)+p_k(i+1)}{2} \quad (4)$$

for : $1 \le i \le I$, $I \le N$

$$a_1=\frac{p_k(i)-p_k(i+1)}{2} \quad (5)$$

for : $1 \le i \le I$, $I \le N$

Part of signal from minimum peak until the consecutive maximum peak can be written as :

$$s_{vp}(n)=a_0+a_1\cos(\omega_1(n)+\pi) \quad (6)$$

for $n = 0,1,2 \ldots N\text{-}1$

Based on equation (3), (4) and (5), the estimates signal equation of the i-th segment between maximum peak until consecutive minimum peak can be written as :

$$s_{pv}(i,n)=\frac{p_k(i)+p_k(i+1)}{2}+\frac{p_k(i)-p_k(i+1)}{2}\cos\left(\frac{2\pi n}{T}+\phi_1(n)\right) \quad (7)$$

for : $1 \le i \le I$, $1 \le N$ and $n = 0,1,2 \ldots N\text{-}1$

Then the period, frequency and phase can be denoted as :

$$T=2.(n_k(i+1)-n_k(i)) \quad (8)$$

$$\omega_1=\frac{\pi}{n_k(i+1)-n_k(i)} \quad (9)$$

$$\phi_1=\frac{\pi.n_k(i)}{n_k(i+1)-n_k(i)} \quad (10)$$

Where $n_k(i)$ is location (sample) of $p_k(i)$. By using the similar value of the period and the phase, we can write the estimated signal from maximum peak until consecutive minimum peak as :

$$s_{pv}(i,n)=\frac{p_k(i)+p_k(i+1)}{2}+\frac{p_k(i)-p_k(i+1)}{2}\cos\left(\frac{\pi.(n-n_k(i))}{p_d(i)}\right) \quad (11)$$

for $p_k(0) > p_k(1)$, then i=0,2,4.....$(I\text{-}2)$ if I is even and
i=0,2,4.....$(I\text{-}1)$ if I is odd
for $p_k(0) < p_k(1)$, then i=1,3,5..., $(I\text{-}1)$ if I is even and
i=1,3,5..., $(I\text{-}2)$ if I is odd

The sinusoidal component coefficients for the estimated signal from minimum peak $p_k(i)$ until the consecutive maximum peak $p_k(i+1)$ are :

$$a_0=\frac{p_k(i)+p_k(i+1)}{2} \quad \text{and} \quad a_1=\frac{p_k(i+1)-p_k(i)}{2} \quad (12)$$

for : $1 \le i \le I$, $I \le N$

The estimated signal for the i-th segment from the minimum until the consecutive maximum peak can be written as :

$$s_{vp}(i,n)=\frac{p_k(i)+p_k(i+1)}{2}+\frac{p_k(i+1)-p_k(i)}{2}\cos\left(\frac{2\pi n}{T}+\phi_1(n)\right) \quad (13)$$

for : $1 \le i \le I$, $I \le N$ and $n = 0,1,2 \ldots N\text{-}1$

By using the period information, frequency and phase used on $s_{pv}(i,n)$, so that the estimated signal from the minimum peak until the consecutive maximum peak can be written as :

$$s_{vp}(i,n)=\frac{p_k(i)+p_k(i+1)}{2}+\frac{p_k(i+1)-p_k(i)}{2}\cos\left(\frac{\pi.(n-n_k(i))}{p_d(i)}\right) \quad (14)$$

If $cos(A) = - cos(A+\pi)$, thus the previous equation can be written as :

$$s_{vp}(i,n)=\frac{p_k(i)+p_k(i+1)}{2}+\frac{p_k(i)-p_k(i+1)}{2}\cos\left(\frac{\pi.(n-n_k(i))}{p_d(i)}+\pi\right) \quad (15)$$

For $p_k(0) > p_k(1)$, then i=1,3,5.... $(I\text{-}1)$ if I is even and
i=1,3,5..., $(I\text{-}2)$ if I is odd
For $p_k(0) < p_k(1)$, then i=0,2,4.....$(I\text{-}2)$ if I is even and
i=0,2,4.....$(I\text{-}1)$ if I is odd

The estimated signal over the frame is a train of the s_{pv}

and s_{vp} for i=0 until i=I-1. based on the previous explanation, for p_k (0) > p_k(1), the reconstructed signal using the segmental sinusoidal model can be written as :

$$s_r(i,n) = \frac{p_k(i) + p_k(i+1)}{2} + \frac{p_k(i) - p_k(i+1)}{2} \cos\left(\frac{\pi.(n - n_k(i))}{p_d(i)} + i.\pi \right) \quad (16)$$

For i = 0,1,2 ...(I-1)

If p_k(0) < p_k(1), the reconstructed signal using the segmental sinusoidal model can be written as :

$$s_r(i,n) = \frac{p_k(i) + p_k(i+1)}{2} + \frac{p_k(i) - p_k(i+1)}{2} \cos\left(\frac{\pi.(n - n_k(i))}{p_d(i)} + (i+1).\pi \right) \quad (17)$$

for i = 0,1,2 ...(I-1)

III. ENCODER

In this paper, a speech signal encoder at 4 kbps and bellow has been designed using several blocks and algorithms. Detail of the encoder is shown in fig. 6. The encoder contains existing signal detector, windowing process, and pitch detector. The next blocks are voiced and unvoiced classificator, sinusoidal based coder, and formant coder.There are some operation mode of the encoder system depends on the kind of signal to obtain the high performance of coding system [10-15]. There are two different operation modes: silent operation mode and signal operation mode. The signal operation mode consists of vibrating mode operation and non-vibrating mode operation. Input signal is speech signal in 16-bit PCM format at 8 kHz frequency sampling. The first block is signal buffer with 30 ms length. The next block is existing signal detector. Then the 30 ms signal will be detected its pitch period width. Based on pitch period information, signal would be classified into vibrating and non-vibrating signal. If it is less than 160 samples, the signal in buffer is called as vibrating signal. Then, if it is more than 160 samples, it is called as non-vibrating signal. The next process is depend on the kind of signal. For vibrating signal (voiced), characteristic signal [16][17] have to be held. One pitch period of signal is quantized using segmental sinusoidal model. The formant information for each pitch period is kept to obtain the variation information changing for 30 ms. The next block is codebook index searching based on periods, peaks, and formants. All of the coded parameter are sent to the decoder with rate 4 kbps or less than 4 kbps, depends on the kind of the speech signal.

The speech input signal exist are detected by using existing signal detector. The signals are buffered with length of 30 ms. The detector identify the input speech signals whether there are signals exist or there are no signals exist. If there are no signals exist, they are called as silence. A sign is transmitted into decoder to inform this condition, so that the decoder is not process the signal during 30 ms. But if there are signals exist, the encoding process is continued with pitch detecting process.

Pitch is the useful parameter in encoding process. Based on the pitch value, we would identify the signals whether voiced or unvoiced. Human speech pitch period of voiced part is vary from 2.5 ms until 20 ms, depend on the gender and the age. Men tend to have the longer pitch period than women and children [18]. The pitch period is detected by using autocorrelation process. The first step, the buffered signal is detected on its peak. Based on the peak value, it can be found the threshold for the center clip process. The threshold is half of the peak value over entire signal in the buffer. The speech signals on the buffer are clipped, so that we would reduce computation complexity. The clipped signals are processed in autocorrelation computation. The autocorrelation process would result two kind patterns. There are peak-valley-peak pattern and peak-valley pattern. The pitch value is detected based on the distance between peaks of the peak-valley-peak pattern. The peak-valley pattern indicates that the signals are unvoiced.

The voiced and the unvoiced signals are classified by using the pitch detection process results. The autocorrelation results pattern are used as reference to identify the signals whether voiced or unvoiced. If the pattern is peak-valley-peak and if the distance between peaks is longer than 2.5 ms but less than 20 ms, it means that the signals is voiced. Then, if the pattern is peak-valley or peak-valley-peak with distance between peaks is longer than 20 ms, it means that the signal is unvoiced. The voiced and unvoiced signals would process in the different methods. The unvoiced signals would be process without referring the pitch period, then the voiced signals would process based on the pitch period.

The voiced signals are fetched on the one pitch period that representing entire voiced signals on the buffer. The one pitch period signals are called as characteristic signal in waveform interpolative signal terminology [16][17]. The length of the characteristics signals is referred as the pitch period. The characteristic signal is quantized on its peaks and periods by using segmental sinusoidal model. For the unvoiced signal, the decimation process is implemented to obtain the smaller size of signal. Then the peak and period quantization is applied. Based on the segmental sinusoidal model, peaks and periods information is extracted. The processed signal would be generated by using the peaks and periods quantization.

The peak information size is reduced by applying 10 look-up tables. The look-up table is also called as codebook. The codebook is trained by using the peak information code-vector. Large amount of the peak information code-vectors are trained with *k-means* algorithm to obtain the peaks codebook. The index number of the peak codebook is varied from 6 to 10 to obtain the optimum process. The period information size is also reduced by applying look-up table. The index number of the peak codebook is also varied from 6 to 10 to obtain the optimum process. The period accuracy has to

maintain to obtain the good receiver perception on the decoder side.

Formant information is important for post-filtering process. The post-filter is arranged as four adaptive band-pass filters to increase the formants and decrease the valley between formants to enhance perception. The formants location is detected by applying the FFT and smoothing filter. Then the low-passed spectra are detected on its peaks location. The peaks location is send to receiver to set the center frequency on the appropriate band-pass filter.

Parameters that have to be sent to the receiver are peaks, periods, pitch, formants, segment, and decimation. The peaks information needs 16-56 bits, the period information needs 6-54 bits, pitch needs 7 bits, segment information needs 6 bits and the formants information needs 0-14 bits. The maximum total coded signal bits resulted for one frame (30 ms) is 120 bits. Thus, the coded speech signals data rate is 4 kbps.

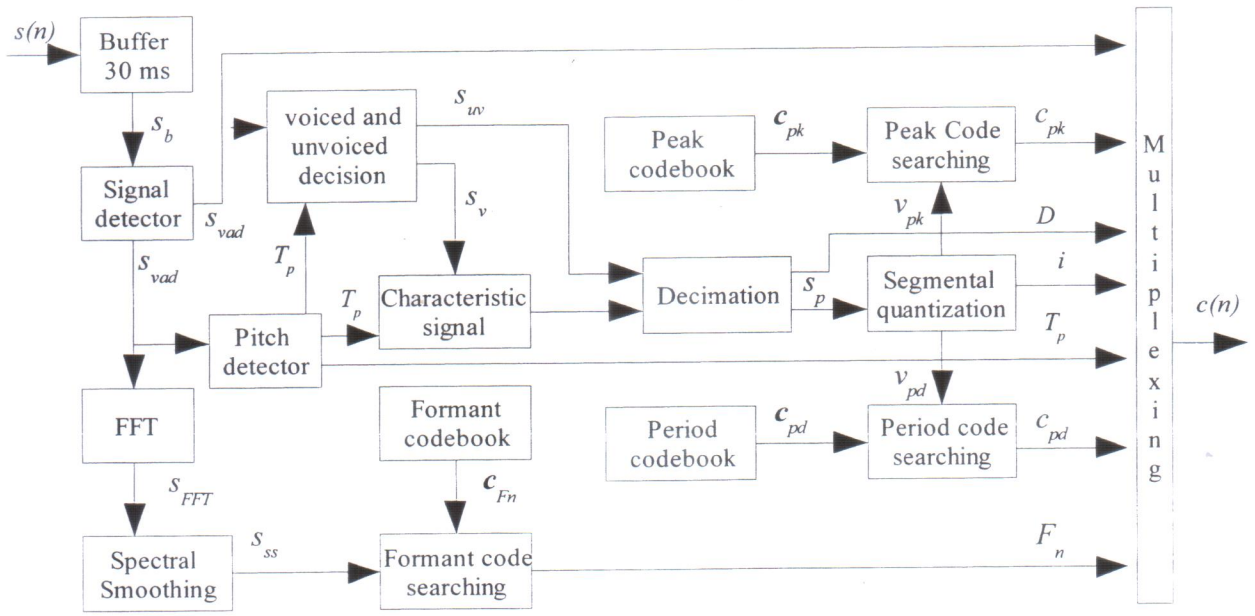


Fig. 6. Block diagram of the encoder

IV. Decoder

In decoder, the coded signal is reconstructed to obtain the speech signal approximation with near toll quality at rate 4 kbps or bellow. If the coded signal is detected as silence, the decoder will generate silence over a frame length i.e. 30 ms. If, it is an existing speech signal, the decoder will generate the speech signal depends on the type of signal.

Unvoiced signal is generated using the sinusoidal parameters. For voiced signal, there is a special processing, that is post-filtering process. The post-filter will enhance performance of the voiced signal by increasing the formant peaks and reducing the valley between formant.

The encoder sends speech signal information as a multiplexed parameter. In the decoder, parameters are de-multiplexed to obtain the useful information for signal reconstruction. Parameters would be generated are peaks, periods information, formants information, pitch, number of segment, and decimation.

The signal detector would identify the kind of signal. If there is a silence, detector would not work because there is no signal on the encoder. If there is a signal, the next process is applied.

The unvoiced signal will be reconstructed if it is detected as unvoiced. The peaks and periods are used to generate the unvoiced signal. The next process is interpolation process with ratio D, inverting of the decimation process on the encoder.

The voiced signal is reconstructed by generating the characteristic signal along the 30 ms segment. The number of characteristic signal along this segment varies between 1.5 until 12, depends on the pitch period. The characteristic signal is generated by using the segmental sinusoidal model from peaks and periods information.

The reconstructed signal, especially the voiced signal is passed into post-filter. The post-filter proposed is a train of four adaptive band-pass filters. The center frequency of each filter is changed every 30 ms segment. The center frequencies are detected by encoder, and then they are sent to decoder. Then the comb filter and compensation filter are applied to improve the hearing perception quality [19].

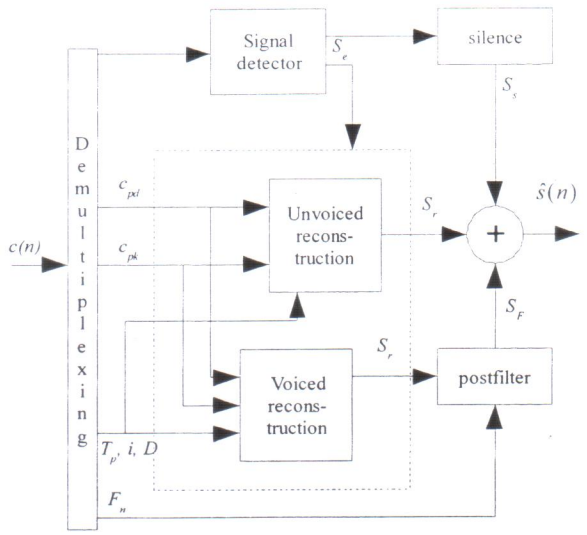


Fig. 7. Block diagram of the decoder

V. Experimental Results

There is a speech signal segment that modeled by sinusoidal approach for each interval between maximum peak to minimum peak and vice versa. This model would result a train of periods which change for every segment with changed amplitude. If amplitude will be kept constant on normalization number, the result was the signal in frequency modulated. Period and peak value for every segment could be used as information which will be sent to the decoder in order to save the transmission channel. In the decoder, the speech signal will be reconstructed to obtain the original signal estimation by sinusoidal model.

Part of signal between minimum to maximum was approached by negative cosine at half period. When the other part of signal between maximum to minimum would be approached by a half period of cosine signal. The number of sinusoidal signals which resulted would vary with respect to the number of peaks on the signal frame. The more peak would decrease compression factor. Fig. 8. shows a part of the original and the reconstructed speech signal from sound "el" taken from word "**el**ektro".

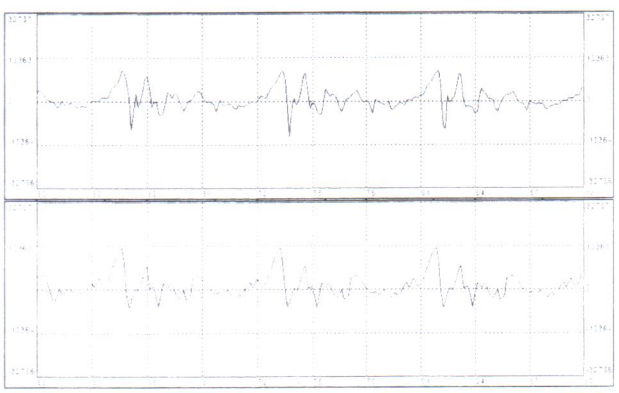


Fig. 8. A 30 ms segment original and reconstructed signal of Indonesian sound "elektro"

The resulted reconstructed signal by sinusoidal approach seems smoother than the original that had arbitrary form between one peak to the consecutive peak. Nevertheless, roughness of original signal means containing high frequency component. Therefore, the spectral power is reduced on the high frequency component. Unfortunately, increasing of the number of peak would decrease the dynamic range of period changing variation for each segment. Thus, compression ratio would be increased to compensate decreasing of compression ratio caused by the number of peaks.

The proposed speech coder has been simulated in a personal computer using a *4kbps coder* program that was specially developed by using C++ programming language. The resulted mean opinion score (MOS) test measured for 15 Indonesian phrase in computer simulation is 3.8 (out of 5). It is tested on 46 peoples with variation on gender, background and age. The coder complexity is less than 15 MIPS, comparable with others kind of speech coder that needs 0.01 MIPS until 90 MIPS. It need less than 30 kB for encoder and 10 kB for decoder.

Then, the coder is implemented on digital signal processor starter kit TMS320VC5416, using a *4 kbps coder* program specially developed for the DSP system, based on *Code Composer Studio ver.2*. Based on the experimental results, hearing perception of the reconstructed signal is fairly good. By using digital signal processor starter kit, the MOS test score is 3.3 (out of 5), because of reducing the number of codebooks. The resulted mean opinion score (MOS) test measured for 15 Indonesian phrase in computer simulation is 3.8 (out of 5). It is tested on 46 peoples with variation on gender, background and age. The coder complexity is less than 15 MIPS. It need less than 16 kB for encoder and 3 kB for decoder. Therefore, it is expected that the proposed low bit rate speech coder with fairly good MOS test score and low complexity will be suitable for voice communication and telemedicine applications during and after disaster cases.

VI. Conclusion

Speech signal could be coded into 4 kbps rate and decoded with high quality of the human perception based on segmental sinusoidal model. The maximum MOS score is 3.8 on Indonesian words. The coder complexity for the digital signal processor implementation is low, it needs less than 10 MIPS.

References

[1] T.F. Quatieri, and R.J. McAulay, "Speech Transformation Based on a Sinusoidal Representation", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34, 1986, pp. 1449-1464.

[2] R.J. McAulay, and T.F. Quatieri,"Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-34,1986, pp. 744-754.

[3] M. Lagrange, S. Marchand, and J.B. Rault, "Sinusoidal Parameter Extraction and Component Selection in A Non Stationary Model", *Proceedings of the 5th International Conference on Digital Audio Effects*,2002, pp. 59-64.

[4] FB Setiawan, S. Tjondronegoro, "Sinusoidal Model of the Speech Signal", Yogyakarta : Proceedings od National Seminar UTY, 2005.

[5] J. Jensen, and J.H.L. Hansen, "A Comparison of Sinusoidal Model Variants for Speech and Audio Representation", *Proceedings of 11th European Signal Processing Conference*, 2004, pp. 479-482.
[6] S. Ahmadi, and A.S. Spanias, "A New Phase Model for Sinusoidal Transform Coding of Speech", *IEEE Transactions on Speech and Audio Processing*, 6, 1998, pp. 495-501.
[7] T. Abe, and M. Honda, "Sinusoidal Model Based On Instantaneous Frequency Attractor", *IEEE Transaction on Speech, Audio and Language Processing*, 14, 2006, pp. 1292-1300.
[8] J. Epps, and W.H. Holmes, Speech Enhancement Using STC-based Bandwidth Extension, *Proceedings of 5th International Conference on Spoken Language Processing*, 1998, pp. 519-522.
[9] F. Ridkosil , *Extreme Waveform Coding*, The Journal of Acoustical Society of America, Vol. 96, issue 4, 1994, pp. 2262.
[10] B.S. Atal, V. Cuperman, and A. Gersho, *Advances in Speech Coding*, Massachusetts: Kluwer Academic Publishers,1991.
[11] J.M. Picket, *The Acoustics of Speech Communication*, Boston : Allyand Bacon, 1999.
[12] L. Rabiner, and B.H. Juang, *Fundamentals of Speech Recognition*, New Jersey : Prentice Hall international,1993.
[13] B.S. Atal, V. Cuperman, and A. Gersho, *Speech and Audio Coding for Wirelles and Network Applications*, Massachusetts : Kluwer Academic Publishers,1993.
[14] D. Jurafsky, and J.H. Martin, *Speech and Language Processing*, New Jersey : Prentice-Hall, 2000.
[15] A.M. Kondoz,: *Digital Speech : Coding for Low Bit Rate Communications Systems*, West Sussex, England : John Wiley & Sons Ltd, 1995.
[16] O. Gottesman, and A. Gersho, "Enhanced Waveform Interpolative Coding at Low Bit-Rate", *IEEE Transactions on Speech and Audio Processing*, 9, 2001, pp. 1-13.
[17] U. Sinervo , *Waveform Interpolation Speech Coding at 2.4-4.0 kb/s*, Master of Science Thesis, Tampere University of Technology, Finland, 2000.
[18] S. Furui, *Digital Speech Processing, Synthesis, and Recognition*, New York : Marcel Dekker Incorporation, 1989.
[19] J.H. Chen, and A. Gersho, "Adaptive Postfiltering for Quality Enhancement of Coded Speech", *IEEE Transactions on Speech and Audio Processing*, 3, 1995, pp. 59-71.

Florentinus Budi Setiawan was born in Semarang (Indonesia) in 1970. He received his undergraduate degree as the best student in Electronics Engineering from the Department of Electrical Engineering, Diponegoro University, Indonesia, in 1993. Since 1994, he joined the Department of Electrical Engineering, Soegijapranata Catholic University, Indonesia as a lecturer. His master (cumlaude) in Telecommunication Engineering was obtained from the Department of Electrical Engineering, Institut Teknologi Bandung, in March 1998, Since 2005, he has been a student in Doctoral Program of the Institut Teknologi Bandung, Indonesia. His current research interests are signal processing, especially on speech coding, disaster mitigation devices development, and railway signaling. **Florentinus Budi Setiawan, ST, MT** is a member of IEEE since 2002.
Contact address : fbudisetiawan@yahoo.com, f_budi_s@unika.ac.id.
Jl. Sinar Pelangi 491, Perum Sinar Waluyo, RT06 RW01, Semarang - 50273, Indonesia.

Author Index

Journal of
eHealth Technology and Application
Published by
Tokai University
NiCT National Institute of Information and Communications Technology

ISSN 1881-4581
9771881458006