Signal Segmentation and Its Application In the Feature Extraction of Speech

Ahmad Idil Abdul Rahman, Sheikh Hussain Shaikh Salleh, Ahmad Zuri Sha'ameri,
Syed Abdul Rahman Al-Attas

Digital Signal Processing Lab
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 Skudai, Johor, Malaysia
Email : ahmadzs@safenet.utm.my

Abstract : Speech is considered as a time-varying signal since the parameters of the signal such as the amplitude, frequency and phase varies in time. Segmenting a duration of captured speech into analysis frames of 20 msecs ensures the assumption of stationarity. If a captured speech segment representing a word that may last for 600 msec, then a total of 30 analysis frames are required to the word. Due to the possibility adjacent frames are identical, then it would be of interest to combine these frames into a single long frame. The interval where adjacent frames have identical parameters is referred as the time-invariant interval (TII). It is of interest to determine these intervals and two methods presented are the instantaneous energy and frequency estimation (IEFE) and localized time correlation (LTC) function. A comparison is made in the accuracy in the TII estimate for a set of speech samples.

Keyworks
Segmentation, feature extraction, speech processing.

I INTRODUCTION

Traditional signal analysis techniques such as spectrum estimation techniques [1] require the signal assumed as time-invariant. In practice, this is not true since signals are time-varying where the parameters of the signal such as the amplitude, frequency, and phase changes with time. For speech processing [2], it is the current practice to segment a captured spoken word into analysis frames of 20 msecs. By doing so, the assumption of time-invariant is approximately applicable to signal within the analysis interval. Only then the parameters of the signal are estimated using spectrum analysis techniques. Examples of spectrum analysis techniques are the periodogram, linear predictive coding (LPC) and cepstrum analysis.

If a spoken word is represented based on fixed length analysis frames, a large number of frames are required to represent the spoken word. For example, the spoken word for the digit one that is 'satu' in Bahasa Melayu can last for a duration of 600 msecs. The total number of frames required to represent the word for an analysis frame of 20 msecs and sampling frequency 8000 Hz is 30 frames. Parts of spoken word can be characterized into unvoice, vowels and plosives. Then, there is a possibility that adjacent analysis frames of the spoken word may have identical parameters since there represent for example the vowel part of the word. It would be of interest to combine these adjacent analysis frames into a single long frame since they represent the same part of the spoken word. The interval where adjacent frames have identical parameters is referred as the time-invariant interval (TII). Based on TII estimate, the number of analysis frames required to represent the word can be significantly reduced. This will aid in minimizing the inputs to a classifier network such as the dynamic time warping (DTW), or artificial neural networks (ANN).

The objective of this paper is to estimate TII, and two methods presented are the instantaneous energy and frequency estimation (IEFE) and localized time correlation (LTC) function. The paper is organized as follows : a model for a time-varying signal is presented, followed by the description of the IEFE and LTC techniques , a comparison in the accuracy for estimating the TII, and the conclusions.

II SIGNAL MODEL

The human speech production apparatus consist of sound source (the glottis) exciting a tube of varying cross-sectional area (the vocal tract). Changing the tube cross-sections over time models the behaviour of the vocal tract. Typical spectrum of human speech is normally contained within the frequency of 300 to 4000 Hz. On the average, the pitch frequency that is due to the vibrating frequency of the vocal cord is 130 Hz for men and 260 Hz for women. Similar to the English language, Bahasa Melayu has phonemes that can be categorize into 3 basic classes:

1. Vowels - a, e, i, o, u
2. Plosives - b, d, p, etc.
3. Fricatives - noise like sounds such as s, z, c etc.

0-7803-6355-8/00/ $10.00 © 2000 IEEE

Within an analysis interval of spoken word, the signal can be expressed as

$$z(n) = \sum_{l=0}^{N_z-1} z_l(n) \qquad 0<n<N\text{-}1 \qquad (1)$$

where N_z is the number of components of the signal, N is duration of the spoken word and $z_l(n)$ is the individual component of the spoken word that represent the phonemes.

Each of the signal component $z_l(n)$ is a limited duration multiple subcomponents pulse sinusoid centered at time n_l samples and of pulse duration N_l samples. The individual signal component do not overlap in time and this ensured by the following conditions

$$n_0<n_1<n_2<\ldots\ldots<n_4$$

$$n_1=n_0+N_0,\ n_2=n_1+N_1,\ \ldots\ldots,\ n_{l+1}=n_l+N_l \qquad (2)$$

If the analytical or complex of signal is assumed, then each individual component of $z_l(n)$ has $N_{l,i}$ subcomponents that is expressed as

$$z_l(n) = \sum_{i=0}^{N_{l,i}-1} \prod_{N_l}(n-n_l)c_{l,i}\exp(j2\pi f_{l,i}(n-n_l)) \qquad (3)$$

where $c_{l,i}$ is the amplitude and $f_{l,i}$ is the frequency of the i-th signal component within $z_l(n)$. The time delay n_l and the pulse duration for each component within $z_l(n)$ is identical since this is in conformance to the condition defined in Equation (2). The term $\prod_{N_l}(n-n_l)$ is referred as a box function that is defined as

$$\prod_{N_l}(n-n_l)=1 \qquad \text{for } n_l-N_l/2<n<n_l+N_l/2$$
$$=0 \qquad \text{elsewhere} \qquad (4)$$

For the spoken word, the number of subcomponents within a phoneme $z_l(n)$ depends on the spoken word and speaker. There is no way to determine this unless actual spectrum analysis is performed on the speech signal. Thus, the number signal component within a phoneme is not easily predictable.

For the male and female speaker the word used is the Malay word 'Lima'. For the simulated waveforms it is define as

Simulated signal 1:

$$x(n) = Ae^{j2\pi f_1 n} \qquad 0<n<N/2$$
$$= Be^{j2\pi f_2 n} \qquad N/2<n<N \qquad (5)$$

Simulated signal 2:

$$x(n) = Ae^{j2\pi f_0 n} \qquad 0<n<N/2$$
$$= B_1e^{j2\pi f_1 n} + B_2e^{j2\pi f_2 n} \qquad N/2<n<N \qquad (6)$$

III ESTIMATING THE TIME INVARIANT INTERVAL OF A SPEECH SIGNAL

A speech signal that represents a spoken word is an example as a time-varying signal that is modelled in Equation (1) to (4). Within an analysis interval, speech signal has intervals where the signal is approximately time invariant. These intervals are referred as the time-invariant interval (TII) and it is of interest in this paper to investigate on methods to estimate these intervals. Instead of segmenting speech into fixed length duration [2], speech is segmented and then its parameters are estimated within the TII. Thus, the number of estimated TII only represents the spoken word where each one has different length compare to the other. Two methods will be presented for used to estimate the TII are the instantaneous frequency estimation (IFE) and the localized time correlation (LTC) function.

A INSTANTANEOUS FREQUENCY METHOD

The concept of instantaneous frequency was discussed in [3] and the methods for estimating the instantaneous frequency was discussed in [4]. Assuming high signal-to-noise ratio conditions, the spoken word can be assumed as

$$z(n) = c(n)\exp(j2\pi\sum_{l=-\infty}^{n} f_i(l)+\phi) \qquad (7)$$

where $c(n)$ is the amplitude, $f_i(n)$ is the instantaneous frequency and ϕ is the phase. Provided the instantaneous energy within a interval of $z(n)$ is nonzero, the instantaneous frequency is

$$f_i(n) = \frac{1}{2\pi}\frac{d}{dn}(\arg[z(n)]) \qquad (8)$$

The instantaneous energy of the signal is

$$E_z(n) = z(n)z^*(n) \qquad (9)$$

The total energy of the signal can be calculated by considering the instantaneous energy over all time. Before the instantaneous frequency is estimated, it is necessary to determine whether a duration of the signal has zero or low energy. If not, instantaneous frequency estimated at this instant will correspond to that of random noise and not of the signal. The instantaneous energy is evaluated for all time instants and compared with a threshold value that is use to

determine whether the instantaneous energy is low. The threshold $E_{z,thd}$ is introduced and is defined as

$$E_{z,thd} = E_{z,\%} E_{z,\max} / 100 \tag{10}$$

where $E_{z,\%}$ is the user defined threshold value in percentage and $E_{z,max}$ is the maximum value of the instantaneous energy.

To minimize the variation in the estimate of the instantaneous frequency, the instantaneous frequency calculated by averaging within an analysis window. The instantaneous frequency estimate is defined as

$$\hat{f}_i(n) = \sum_{l=-N/2}^{N/2} g(l) f_i(n-l) \tag{11}$$

where $g(n)$ is the analysis window. The analysis window is

$$g(n) = \frac{1}{N_g} \qquad -\frac{N_g}{2} < n < \frac{N_g}{2} \qquad N_g << N \tag{12}$$

where N_g is the analysis window width which is fixed and is independent of time.

The time instant λ_1 is the lower TII and is chosen such that the instantaneous energy is nonzero. If the instantaneous frequency at λ_1 is constant, the its estimated for all time intervals upto time instant λ_2.is also constant. This is only true if the time interval $\lambda_1 < n < \lambda_2$ is the true TII where λ_2 is less N. If time instant λ_1 is defined, the location of λ_2 can be estimated by evaluating the inequality

$$\hat{f}_i(\lambda_1) \neq \hat{f}_i(\lambda) \pm f_{i,\text{limit}}, \lambda \in t, 0 < \lambda < N-1 \tag{13}$$

where f_{limit} is the predetermined acceptable variation in $\hat{f}_i(\lambda_1)$. As λ is incremented over the time axis, the value of λ that satisfies the inequality is upper limit of TII. Once the upper limit of the present TII λ_2 is determined, the next step is to determine the next TII of the spoken word. The time instant λ_2 is used as the lower limit of the next TII and the same procedure based on Equation (11) is repeated again to estimate the upper limit of the TII. For the signals used for analysis, the user defined threshold value $E_{z,\%}$ used is 0.25, and the acceptable frequency variation f_{limit} is 100 Hz.

B LOCAL TIME CORRELATION METHOD

The correlation function [5] measures the similarity between signals and is defined as

$$R_{z_1,z_2}(\lambda_1,\lambda_2) = \sum_{n=0}^{N-1} z_1(n-\lambda_1) z^*_2(n-\lambda_2) \tag{14}$$

$R_{z_1,z_2}(\lambda_1,\lambda_2)$ is known as the crosscorrelation function if $z_1(n)$ and $z_2(n$ are not the same function. If $z_1(n)$ and $z_2(n)$ are the same, $R_{z_1,z_2}(\lambda_1,\lambda_2)$ is the autocorrelation function. In this application, it is desired to measure the correlation between $z_1(n)$ and $z_2(n)$ for a given time frame at an instant in time. Thus, a special form of correlation function known as the local time correlation function (LTC) is

$$R_{loc,z_1,z_2}(\lambda_1,\lambda_2) = \sum_{n=0}^{N-1} z_{loc,1}(n-\lambda_1) z^*_{loc,2}(n-\lambda_2) \tag{15}$$

where $z_{loc,1}(t)$ is a segment in time of signal $z_1(n)$, $z_{loc,2}(n)$ is a segment in time of signal $z_2(n)$ and λ_1 and λ_2 is the time instant of interest. The localized functions $z_{loc,1}(n)$ and $z_{loc,2}(n)$ are defined as

$$\begin{aligned} z_{loc,1}(n) &= g(n) z_1(n) \\ z_{loc,2}(n) &= g(n) z_2(n) \end{aligned} \tag{16}$$

where $g(n)$ is the analysis window function similar to that defined in Equation (12). The width of the analysis window is constrained by the period of the signal of interest. If the analysis window width is made smaller compared to the period of the signal, then the characteristics of the signal is not represented within the analysis window. For example, the typical spectrum of speech ranges from 300 to 4000 Hz. Thus, analysis window of width greater than 32 samples or are sufficient to include at least two cycles of the signal component with the lowest frequency.

Based on the definitions of the $z_{loc,1}(n)$ and $z_{loc,2}(n)$, the localized time correlation (LTC) function can be expressed as

$$R_{loc,z_1,z_2}(\lambda_1,\lambda_2) = \sum_{n=0}^{N-1} |g(n)|^2 z_1(n-\lambda_1) z^*_2(n-\lambda_2) \tag{17}$$

The average localized energy of the signal at instant λ is

$$E_{loc,z_1}(\lambda) = \sum_{n=0}^{N-1} [g(n) z_1(n-\lambda)]^2 \tag{18}$$

If normalized to the square-root of the total localized energy, the normalized LTC function is

$$R_{z_1,z_2}(\lambda_1,\lambda_2) = \frac{R_{loc,z_1,z_2}(\lambda_1,\lambda_2)}{\sqrt{E_{loc,z_1}(\lambda_1) E_{loc,z_2}(\lambda_2)}} \tag{19}$$

The range of possible values of the normalized LTC function is

$$0 < \left| R_{z_1,z_2}(\lambda_1,\lambda_2) \right| < 1 \quad for \lambda_1,\lambda_2 \in n \tag{20}$$

A value of 1 indicates the highest correlation between $z_1(n)$ and $z_2(n)$ for a given instant λ while the lowest correlation will be presented by a zero value.

If comparison is to be made on the same signal $z(n)$ but at different time instants, then the LTC function, localized energy and the normalized LTC function are defined as

$$R_{loc,z}(\lambda_1,\lambda_2) = \sum_{n=0}^{N-1} \left| g(n) \right|^2 z(n-\lambda_1) z^*(n-\lambda_2) \tag{21}$$

$$E_z(\lambda_1) = \sum_{n=0}^{N-1} \left[g(n) z(n-\lambda_1) \right]^2 \tag{22}$$

$$R_z(\lambda_1,\lambda_2) = \frac{R_{loc,z}(\lambda_1,\lambda_2)}{\sqrt{E_{loc,z}(\lambda_1) E_{loc,z}(\lambda_2)}} \tag{23}$$

The normalized LTC function is more convenient to use compared to the LTC function since the magnitude of the function is independent of the amplitude of the signal. For example, two finite duration complex sinusoids are defined as

$$z_1(t) = Ae^{j2\pi f_1 n}$$
$$z_2(t) = Be^{j2\pi f_1 n} \qquad \text{for } -N/2 < n < N/2 \tag{24}$$

The LTC and normalized LTC function of the signals are

$$R_{loc,z_1z_2}(0,\lambda) = ABe^{j2\pi f_1 \lambda}$$
$$R_{z_1z_2}(0,\lambda) = e^{j2\pi f_1 \lambda} \tag{25}$$

The example shows that the two signals are correlated to each other at all values of time instant λ and the highest correlation occurs at λ=0. The normalized LTC function is independent of the amplitude of the signal and no knowledge is required on the amplitude or the energy of the signal.

If the instantaneous energy of the time-varying signal is nonzero, the time instant λ_1 is used as the lower limit of the TII and the LTC function has a finite value at this instant. The LTC function is constant upto the upper limit of the TII at time instant λ_2. The TII is then defined as the interval from λ_1 to λ_2. If the lower limit of the TII is defined at time instant λ_1, the location of λ_2 can be estimated by evaluating the inequality

$$\left| \mathcal{R}_z(\lambda_1,\lambda) \right| < R_{z,thd} \left| \mathcal{R}_z(\lambda_1,\lambda_1) \right| \tag{26}$$

where $R_{z,thd}$ is the correlation threshold level used to decide if the upper limit of the TII has been reached. If the inequality is satisfied, then λ is the upper limit of the local stationary interval.

Once the upper limit of the local stationary interval λ_2 is determined, the next step is to determine the next TII of the spoken word. The time instant λ_2 is now used as the lower limit of the TII and the same procedure is repeated again to estimate the upper limit of the TII. Before the LTC method is used, it is necessary to check the signal energy level using the procedure described in equation (10). The correlation threshold of 0.5 can be used provided that the amplitude of the signal at adjacent TII is similar. However, this is not true since the amplitude of speech is time-varying and is not equal from one component to another. The estimated TII may not correspond to the true TII. Further experimentation is required to determine the suitable range of value in the correlation threshold value so that the range of estimated TII is approximately the true TII.

IV RESULTS

The objective of this paper is to find the method that can accurately estimate the TII of the set of test signals. For the simulated signals, both the IFE and LTC shown in Figure 1 and 3 can accurately estimate the TII.

The spoken word for 'lima' is shown in the time-frequency representation in Figure 4. For both male and female speakers, the TII shown in Figure 5 and 6 for the spoken word can be represented using the LTC. However, TII cannot be represented by the IFE method as shown in Figure 7. This is because of the rapid variations in the instantaneous energy of the spoken word. The duration of low signal-to-noise ratio conditions that increases the variance of the instantaneous frequency estimate. Thus, the LTC is the more appropriate method for speech applications.

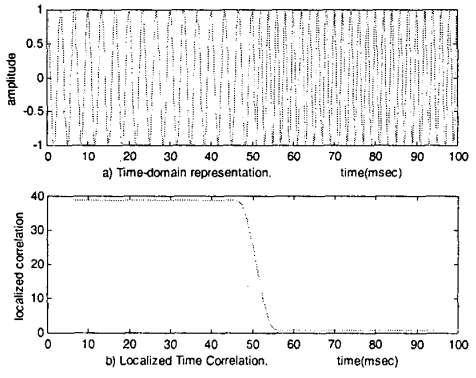

Fig. 1 Time domain and LTC representation for the simulated signal 1.

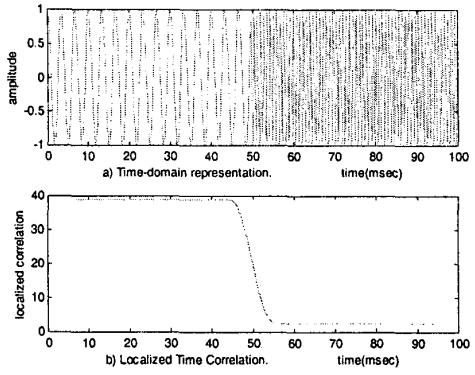

Fig. 2 Time domain and LTC representation for the simulated signal 2.

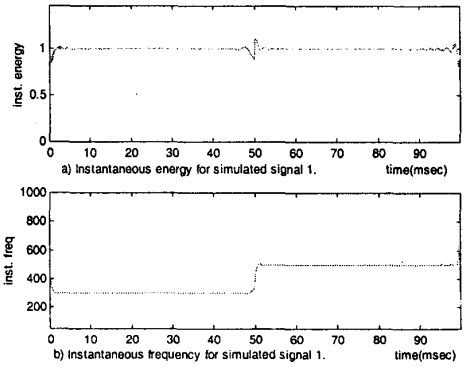

Fig. 3 Instantaneous energy and instantaneous frequency representation for the simulated signal 1.

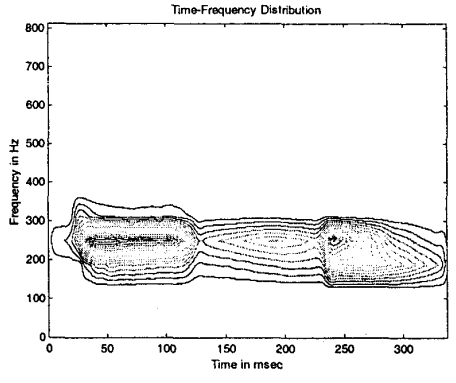

Fig. 4 Time-frequency representation for the spoken word 'lima'.

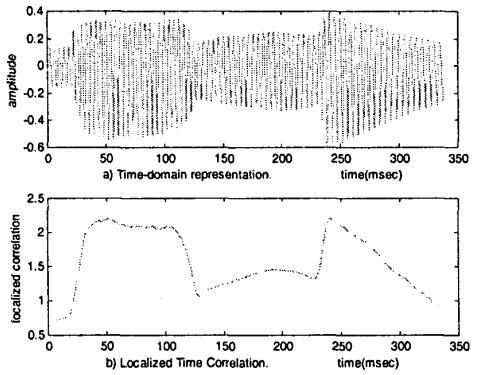

Fig. 5 Time domain and LTC representation for the spoken word 'lima' from female speaker.

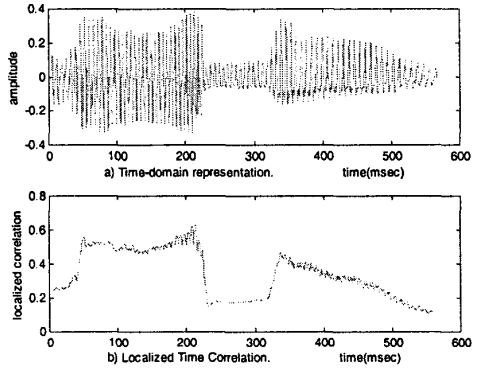

Fig. 6 Time domain and LTC representation for the spoken word 'lima' from male speaker.

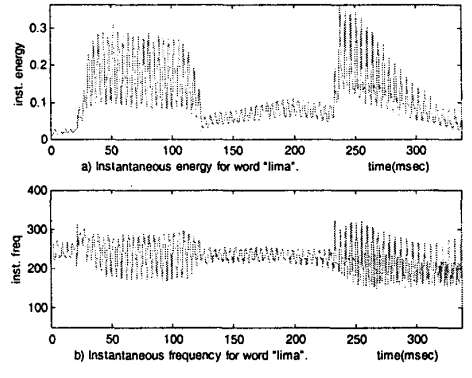

Fig. 7 Instantaneous energy and instantaneous frequency representation for the spoken word 'lima'

V CONCLUSIONS

Signal such as spoken word is considered as a time-varying since the parameters of the signal such as the amplitude, frequency and phase varies in time. To represent this signal, fixed length frames are applied to ensure that the signal is approximately time-invariant within the frame. The total number of frames can be reduced since there is a possibility that adjacent frames are similar. These intervals are known as the TII and the parameters of speech are represented within this interval. The IFE and LTC function are introduced to estimate the TII and the total number of frames required to represent spoken word is reduced. For speech signals, it is found that the LTC function is a more appropriate method to estimate the TII. Thus, the number of frames using this method depends only on the phonemes present for a given spoken word.

VI REFERENCES

[1] Kay, S. M., *Modern Spectral Estimation*, Prentice-Hall, New Jersey, 1988.

[2] Schaffer,R.W., Rabiner,L.R., *Digital Processing of Speech Signal*, Prentice Hall, New Jersey, 1978.

[3] Boashash, B., "Estimating & Interpreting the Instantaneous Frequency of a Signal-Part I: Fundamentals", *Proc. of the IEEE*, Vol.80, No.4, April 1992, pp. 519-538.

[4] Boashash, B., "Estimating & Interpreting the Instantaneous Frequency of a Signal-Part 2: Algorithms and Applications", *Proc. of the IEEE*, Vol 80, No 4, April 92, pp540-568.

[5] Peebles, P.Z., *Probability, Random Variables and Random Signal Principles*, McGraw-Hill, New York, 1987.