SPEECH PROCESSING IN DIGITAL SIGNAL PROCESSING

N. H. M. Safee¹, N. M. Z. Hashim², M. Daud³, A. M. Darsono⁴, A. Salleh⁵, N. R. Mohamad⁶
¹-⁶(Faculty of Electronic and Computer Engineering
UniversitiTeknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, Melaka, Malaysia)

ABSTRACT
Digital signal processing is concerned with both a discrete signal representation, and with the theory, design and implementation of numerical procedures for processing discrete representation. Techniques of digital signal processing that does not mean special consideration in the context of speech communication. There are several ways characteristic and its algorithm of speech processing. The DSP methods are used in speech analysis, synthesis, coding, recognition, enhancement as well as voice modification, speaker recognition, language identification.

Keywords–Frequency, Processing, Recognition, Signal, Speech, System

I. INTRODUCTION
Speech is one of the ancient ways to specify ourselves. Today the speech signals are also used in biometric recognition technology and communicate with the machine. The speech signal is slowly timed different signal (quasi-stationary). When examined in a short-lived (5-100 msec), features a relatively stationary. But, if for the time change characteristics of the signal, it reflects the different speech sounds spoken. The information in the speech signal actually is represented by the short-term spectral amplitude waveform speech (phonemes).

In speech communication systems, the speech signal are transmitted stored and processed in many ways. Technical concerns lead to a wide variety of representations of the speech signal. In general, there are two major concerns in any system:

(i) Preservation of the message content in the speech signal
(ii) Representation of the speech signal in a form that is convenient for transmission or storage or in a form that is flexible so that modification may be to the speech signal without seriously degrading the message content.

II. APPLICATION OF SPEECH PROCESSING
There are some applications that use in Speech Processing:

A. Speech Production Models and Digital Implementation
In this application it will introduce the basic physical phenomena involved in speech production and digital implementations of these models.

The speech is a produced by acoustically exciting a time varying cavity such as the vocal tract. The vocal tract which is the mouth cavity bounded by the vocal cords and the lips. The speech sounds are produced by adjusting both the type of excitation as well as the shape of the vocal tract. There is several of classifying speech sounds:

- Voiced – The sound are produced by the tract by quasi-periodic puffs of air produced by the vibration of the level cords in the larynx. By the vibrating cords it will be modulate the air stream from the lungs at a rate may be as low as 60 times per second for some male to as high as 400 or 500 times per second for children.
- Nasal - A part or all of the airflow is diverted into the nasal tract by opening the velum.
- Plosive – The sound are produced by exciting the tract by a sudden release of pressure
- Fricatives – Produced by turbulent flow created by air flow through a narrow constriction.
- Voice fricatives – The tract simultaneously by turbulent and by vocal cord vibration.
- Affricates – The sound that begin as a stop and are released as a fricative.

In the digital implement, the standard theory of sampling in the time and frequency domains is used to
convert the continuous signals considered to sample signal and the sample represented digitally to the desired number of bits per sample. There are two sets specification of parameter are needed the parameters that specify the shape of the vocal tract and those that control the glottis. The vocal tract parameters implicitly control nasality and also frication.

B. Speech Coding

The application of digital speech processing technology involves storage and transmission. It is because the goal is to compress the digital representation of speech wave representation into low bit rate. It is called "speech coding". The bit rate and speech-related quality is dependent on applications such as at low bit rate. It is basically the rate of between 75 and 2400 bps (bits per second) and moderate to high bit rates it will operate at more than 2400 bps.

Speech coders allows variety of applications including narrowband and broadband wired telephony, cellular communications, voice over internet protocol (VoIP) and secure voice for privacy and encryption. The Speech coders frequently use in speech production and speech perception process, and therefore might not be useful to many general audio signals such as music.

C. Text-to-Speech Synthesis

Synthesis of Speech is the reverse process to the recognition. It also known generated the speech signal using computational means for effective human machine interactions. Text-to-phoneme conversion is a once the synthesis processor has determined the set of words to be spoken, it must derive pronunciations for each word. Word pronunciations may be conveniently described as sequences of phonemes, which are units of sound in a language that serve to distinguish one word from another.

D. Speech Recognition

Speech recognition is a technology that allows the control of machines by voice. It involves there cognition and understanding of the language spoken by the machine. Speech recognition technology is based on the pattern recognition. The objective is to take the input patterns; the speech signal and classifies it as a sequence of pattern has defined the right saved. The stored patterns can be made from the unit, which is calling the phonemes. If the pattern is unchanged and will not change the speech, there will be no problem but just compares the sequence features with the stored patterns, and find the right match when it happens. However basic difficulty is the speech recognition the speech signal is

---

**Fig. 1** Speech coding block diagram encoder and decoder

**Fig. 2** Text-to-speech synthesis system block diagram
highly variable due to different speakers, different speech rates, of different content and the different acoustic conditions. The task is to determine which of the variety in the speech recognition and speech related to variations are not relevant.

The characterization of speech recognition system:
(i) The manner in which a user speaks to the machine.
   There are generally three modes of speaking:
   • Isolated word mode in which the user speaker individual words drawn from a specified vocabulary.
   • Continuous speech mode in which the user can speak fluently from a large vocabulary.
   • Connected word mode which the user speaks fluent speech consisting entirely of word from a specified vocabulary.

(ii) The knowledge of the user’s speech pattern, including:
   • Speaker dependent systems which have been custom tailored to each individual talker.
   • Speaker independent system which works on broad population of talkers means the system never encountered or adapted.
   • Speaker adaptive system which customize their knowledge to each individual user over time while the system is in use.

(iii) The amount of acoustic and lexical knowledge use in the system:
   • Simple acoustic system which have no linguistic knowledge.
   • System which acoustic and linguistic knowledge, where is generally represented via syntactical and semantic constraints on the output of the recognition system.

(iv) The dialogue between the human and the machine:
   • One-way communication in which each user spoken input is acted upon.
   • System-driven dialog system in which the system is the sole initiator of a dialog requesting information from the user via verbal input.
   • Natural dialogue system in which the system machine conducts a conversation with the speaker, solicits inputs.

III. METHODOLOGY OF SPEECH PROCESSING

Linear Predictive Coding (LPC) is single of the most influential speech analysis techniques. It is valuable method for encoding quality speech at a low bit rate. The basic idea behind linear predictive analysis is that a specific speech model at the current time can be projected as linear combination of past speech model.

Human speech is created by a LP as a model base. Conventional source filter model is utilized, in which the glotta, vocal tract and lip radiation transfer functions are combined into one all-pole filter that simulates audibility of the vocal tract. The principle behind the use of the LPC is to reduce the quantity of the squared differences between the original speech signal and the approximated speech signal over a finite duration. This can be used to provide a unique set of predictor coefficients. Every frame is estimated by these predictor coefficients, which is normally 20 ms long. Ak is represented of predictor coefficient. The gain (G) is another important parameter. The transfer function of the time varying digital filter is given by:

\[ H(z) = \frac{G}{1 - \sum akz^{-k}} \]  

Where k=1 to p, which will for the LPC-10 algorithm and 18 for enhanced algorithm that is used. Levinson-Durbin recursion will be used to compute the necessary parameters for the auto-correlation method.

The LPC analysis of each frame also contains the decision-making process of voiced or unvoiced. A pitch-detecting algorithm is employed to define to correct pitch period or frequency. It is significant to re-emphasis that the pitch, gain and coefficient parameters will be fluctuating with period from any frame to another.

E. Performance Analysis

Parameters are involved in performance evaluation of LPC’s:
• Bit Rates
• Overall Delay of the System
• Computational Complexity
• Objective Performance Evaluation

F. Types of LPC

The categories of LPC are following:
• Voice-excitation LPC
• Residual Excitation LPC
• Pitch Excitation LPC (MPLPC)
• Regular Pulse Excited LPC (RPLP)
• Coded Excited LPC (CELP).

IV. THE ALGORITHM FOR SPEECH SIGNAL ANALYSIS

The Fourier-Bessel expansion is to isolate multiple formants of speech signal in one step and not needful previous information about frequency band of
the resonance signals. Below is steps described the algorithm proposed for speech signal analysis:

1. Filtering

Fourier-Bessel (FB) is worked to separate speech resonances. Zero order FB series expansion of band limited signal $r(t)$ over time interval $(0,T)$ is defined as:

$$x(t) = \sum_{l=1}^{Q} C_l J(\frac{\lambda_l t}{T})$$

Where $\lambda_l, l = 1,2,3,... Q$ cumulative orders are positive roots for $J_0(\lambda_l) = 0$, and $J_0(\frac{\lambda_l t}{T})$ are zero-order Bessel functions. By using orthogonality of zero order Bessel function FB coefficient $C_l$ can be designed as:

$$C_l = \frac{2}{\pi [J_1(\lambda_l)]^2} \int_0^T x(t) J_0(\frac{\lambda_l t}{T}) dt$$

For $l = 1,2,..,Q$ where $J_1(\lambda_l)$ is first order Bessel function. By abuse orthogonality of zero order Bessel functions, FB coefficient $C_l$ can be desined. It has been exposed that the order of non-zero coefficients and FB series expansion of the adapted test signal as the center frequency and bandwidth of the signal differ. It is expected that the function is integral over interval. It is prove that the order and range of nonzero coefficient of the FB series expansion of a test signal are altered as the center frequency and the bandwidth of the signal are diverse. It is shown that range widens with larger bandwidth and order ascending with higher center frequency. Explanation of above can be demonstrate analytically for the signal

$$x[n] = \cos(\omega n)$$

Which yields:

$$C_l = \frac{2\lambda_m \cos[\omega(n-\alpha)]}{\lambda_m \pi^2 + \omega^2 [N^2]}$$

Where:

$$\alpha = \sin^{-1}(\frac{\omega N}{[\lambda_m^2 - \omega^2 N^2]^2 + \omega^2 N^2])$$

The peak value of $C_l$ is attained for order $l$ where the root $\lambda_m = \omega N$ and value $C_l$ is given by

$$C_{l,peak} \approx \frac{2 \sin(\omega N)}{\lambda_m \omega N}$$

The magnitude of $C_l$ reductions quickly from order where the peak value is occurring and the value is unimportant at far away orders. It is clear from above there is one to one correspondence between frequency content of a signal and the order of FB expansion where the coefficient attains peak magnitude. FB expansion signifies the spectrum better than Fourier expansion for each of the basic functions of Bessel support limited bandwidth around the center frequency. FB expansion was found to be suitable to reduce the cross-term in the Wigner-Ville distribution and the immediate detection of glottal closure / time (GCIs) of the speech signal. In the proposed method, it is expected that a good speech is separated in the frequency domain, and the resonance will be related with a variety of different non-overlapping clusters FB coefficients. Thus, each formant speech can be rebuilt separately by identifying whether the FB coefficients. Time interval $T$ should be properly limited to the immediate detection of formants frequencies and amplitudes vary rapidly with time does not cause problems.

2. Formant Analyses

Formant analyses by a Hilbert transform algorithm separation: HTSA is used to get the function AE and IF functions. Below is symbolized the single speech resonance:

$$x[n] = \alpha[n] \cos(\Phi[n])$$

The instantaneous frequency and amplitude envelope of speech resonance are estimated by Hilbert transform separation algorithm as

$$x_H[n] = \alpha[n] \sin(\Phi[n])$$

$$x[n] + jxH[n] = \alpha[n] e^{j\Phi[n]}$$

$$\alpha[n] = \sqrt{x^2[n] + x^2H[n]}$$

$$\Phi[n] = \tan^{-1}(\frac{xH[n]}{x[n]})$$

Ripples on show in the amplitude envelope function a (n) and instantaneous frequency function $\Phi(n)$ are smoothed by using filters.

3. The instantaneous frequency (IF)

The instantaneous frequency (IF) and time varying Amplitude is estimated: DESA considers individual sample based signal analysis is affected due to presence of noise in the speech signal. The Hilbert transform is applied on separated formants to estimate the instantaneous frequency (IF) and time amplitude (AE) of the speech formant. The instantaneous parameters are functions of time and can be estimated at any point of the signal the total number of points that map the signal is much better than the number of peak points of the signal. Thus, the Hilbert transform separation algorithm (HTSA) provides the best available time resolution of the amplitude and other estimations signal analysis. It also paved the way for the average and other statistical processing procedures make the analysis more accurate signal.

The frequency resolution means the smallest change or granularity with one can differentiate details in the frequency domain of a signal. The spectral resolution is a discrete and terminal value that cannot be less than the ratio between the sampling frequency and the number of measured point for any other digital method. Instead, the HTSA calculates the frequency as

www.ijaert.org
the rate of phase change with time, and the instantaneous frequency resolution can be as small as desired depending on the accuracy of the signal phase and time measurement.

Practically, the HTSA is suggesting computational algorithms are simple and fast thus allowing online (real-time) signal analysis and identification. Therefore, the HTSA methods are suitable for analysis and identification of both linear and nonlinear systems; it is produce adaptive decomposition of complicated multi component signals. With the help of the HTSA, the researchers get new features such as the envelope and the instantaneous frequency of each component and the instantaneous phase relations between time-varying components.

V. CONCLUSION

This paper is discussed about the Speech Processing in Digital Signal Processing. There are applications such as Speech Production Models and Digital Implementation, Speech Coding, Text-to-Speech Synthesis and Speech Recognition. Also, this paper discussed about the methodology of Speech Processing. Actually, many kind of the methodology of speech processing signal but LPC model is chosen to discuss. From the speech signal analysis it use FB expansion to separate multiple speech formants in one step and do not requires prior information of frequency band of speech formants. The HTSA pair provides a method for determining the instantaneous frequency and amplitude of a signal. The method is also good for solving problems concerning analysis of stationary and non-stationary signals, as well as narrow-and/or wide band multi-component signals.

VI. ACKNOWLEDGMENT

We are grateful to Centre for Telecommunication Research and Innovation (CeTRI) and UniversitiTeknical Malaysia Melaka (UTeM) through PJP/2013/FKEKK (29C)/SO1215 for their kind and help for supporting financially and supplying the electronic components and giving their laboratory facility to complete this study.

REFERENCES


