AC 2012-5055: MULTIMEDIA SYSTEMS EDUCATION INNOVATIONS I: SPEECH

Prof. Tokunbo Ogunfunmi, Santa Clara University

Tokunbo Ogunfunmi is the Associate Dean for Research and Faculty Development in the School of Engineering at Santa Clara University (SCU), Santa Clara, Calif. He is also an Associate Professor in the Department of Electrical Engineering and Director of the Signal Processing Research Lab. (SPRL). In 2003, he served as Acting Chair of the Department of Electrical Engineering at SCU. His research interests include digital signal processing, adaptive and nonlinear filters, multimedia (Video/Audio/Speech), neural networks, and VLSI/FPGA/DSP development. He has published 140+ papers in refereed journal and conference papers in these and related areas. He has also published two books: Adaptive Nonlinear System Identification: Volterra and Wiener Model Approaches, published by Springer, 2007, and Principles of Speech Coding, co-authored with Dr. Madihally (Sim) Narasimha and published by CRC Press, 2010. He is a Senior Member of the Institution of Electrical and Electronic Engineers (IEEE), a member of Sigma Xi (the Scientific Research Society), American Association for Engineering Education (ASEE), and a member of the American Association for the Advancement of Science (AAAS). He is on the editorial board of IEEE Signal Processing Letters and Circuits, Systems, and Signal Processing. He is the Chair of the IEEE CASS Technical Committee on Circuits and Systems for Education Outreach (CASEO). He obtained his B.S.E.E. (first class honors) from University of Ife, Nigeria, and M.S. and Ph.D. in electrical engineering from Stanford University.

Multimedia Systems Education Innovations: Part 1(Speech)

Abstract

Multimedia Systems is becoming very important in undergraduate education.
The word multimedia refers to speech, audio and video data.
Speech, audio and video and general digital signal processing (DSP) devices are very common-place in everyday life. This is due to the growth of popularity of personal digital assistants (PDA), cellular phones, and other embedded speech/audio/video devices. One of the major applications of DSP processors is in speech, audio and video processing.

We are developing a 3-part course sequence that will help teach undergraduates multimedia systems. The first part of the 3-part course sequence is about speech. The second and third parts will focus on audio and video. The goal is to demystify these applications of real-time multimedia signal processing so that undergraduates can have a mastery of the basic techniques.

In this paper, we provide details of a course we designed for undergraduates which focuses on understanding how to process speech. We provide examples of the curriculum, what is covered and how we cover it. We also provide examples of laboratory projects that are used to complement the class lecture sessions. We use MATLAB software in all the lab projects.

We also discuss possible implementations of the speech coding and processing using hardware such as DSPs. In the future, we plan to introduce the use of FPGAs for this application as well.

Details of the course and our experiences in developing and offering them will be presented at the conference.

Introduction

Previously, we have developed three graduate-level courses in the Multimedia area of Speech to teach the fundamentals of speech coding and voice-over-IP. They are a 3-course sequence (1) ELEN 421 (Speech Coding I) (2) ELEN 422 (Speech Coding II) and (3) ELEN 423 (Voice-over-IP). The focus of the courses are on the Advanced DSP courses in the area of Speech. More details about these courses are in [1]. We discussed the pedagogy (the principles and methods of instruction) for our follow-up graduate courses on Speech Processing and Voice-over-IP. We focused on the activities of educating or instructing; activities that impart knowledge or skill in these areas.
We have also recently published a textbook that contains some of the material taught in the first two courses [2].

In this paper, we provide details of a course (proposed ELEN 135 Multimedia Systems I: Speech) we designed for undergraduates which focuses on understanding how to process speech signals. We provide examples of the curriculum, what is covered and how we cover it. We also provide examples of laboratory projects that are used to complement the class lecture sessions. We use MATLAB software in all the lab projects.

We also discuss possible implementations of the speech coding and processing using hardware such as DSPs. In the future, we plan to introduce the use of FPGAs for this application as well.

The paper is divided into five sections. In Section 2, we start with the discussion of Systems and DSP curriculum improvement. In Section 3, we give some details of the Multimedia Systems I: Speech course. In Section 4, we present a sample course project on implementation of an encoder of the Split-Band LPC speech coder.
In Section 5, we present a SIMULINK implementation and conclude in Section 6.
We hope that our experiences maybe useful for other faculty considering an undergraduate course in multimedia systems for speech. Future work will report on our proposed development of multimedia systems for audio and video.

Systems and DSP Curriculum Improvement

Many universities, including ours, continually strive to improve their programs by assessing its impact and learning outcomes and modifying, changing or deleting, adding courses based on academic and industrial technology trends. This is actually required by the Accreditation Board for Engineering Technology (ABET) [3] as part of accreditation requirements. In the area of Digital Signal Processing (DSP), many schools offer a single course introducing the theoretical methods used. What is lacking in most cases, is a course on an application area of DSP such as multimedia (speech, video and audio) systems.

We have proposed a new course (ELEN 135 Multimedia Systems I: Speech) designed for undergraduates which combines salient topics in the two graduate courses with appropriate labs and projects.

Details of the Multimedia Systems I: Speech Course

ELEN 135 Multimedia Systems I: Speech

Review of sampling and quantization. Introduction to Digital Speech Processing. Elementary principles and applications of speech analysis, synthesis, and coding. Speech signal analysis and modeling. The LPC Model. LPC Parameter quantization using Line Spectrum Pairs (LSPs). Digital coding techniques: Quantization, Waveform coding. Predictive coding, Transform coding, Hybrid coding and Sub-band coding. Applications of speech coding in various systems. Standards for speech coding.

Advanced aspects of speech analysis and coding. Analysis-by-Synthesis (AbS) coding of speech, Analysis-as-Synthesis (AaS) coding of speech. Code-Excited Linear Speech Coding. Error-control in speech transmission. Application of coders in various systems (such as wireless phones). International Standards for Speech Coding. Real-Time DSP

implementation of speech coders. Research project on speech coding. Introduction to speech recognition.
Prerequisite: ELEN 110 (Linear Systems) and ELEN 133 (DSP) or equivalent. (2 units)

The textbook for the course is [2]. The aim of the course is to understand the introductory methods of speech coding. Other books and papers [4-8] are used as reference.

We enable understanding the advanced methods used speech coding. Concepts such as vector quantization, analysis-by-synthesis methods used in CELP coders are introduced and applied. Many examples are given of the concepts as specified in various of the standardized speech coders from ITU and other agencies. In particular, we use the FS 1015 Speech Coder, the G.729A Speech Coder and the internet Low Bit Rate codec (iLBC) as examples.

Lab Contents

The laboratory part of the course consists of ten possible laboratory exercises.
Some of the labs require the use of MATLAB. The titles of the labs are :
Lab 1: Review of DSP Methods and Techniques for Speech
Lab 2: Pitch Determination Methods
Lab 3: LPC Model. LPC Parameter quantization using Line Spectrum Pairs (LSPs)
Lab 4: Scalar Quantization: Nonlinear Quantizers in Speech (mu-law and A-law)
Lab 5: Example: G.711 Speech Coder
Lab 6: Vector Quantization
Lab 7: Analysis by Synthesis Methods: CELP Coders
Lab 8: Example: The FS 1015 Speech Coder or The G.729A Speech Coder
Lab 9: Example: The internet Low Bit rate Codec (iLBC)
Lab 10: MATLAB / SIMULINK / DSP implementation of speech coders

These labs build on one another and mid-point of the course leads to the student implementing in MATLAB the standardized ITU G.711 voice coder [7] which is commonly used in industrial applications. These rest of the labs also build on one another and culminates in the student implementing in MATLAB examples speech coders such ad the standardized ITU G.729A voice coder [8] which is and commonly used in industry. As advances are made in speech coding technologies, we have added new topics to the course contents. One example is the discussion and evaluation of the internet Low Bit Rate codec (iLBC) recently proposed and beginning to be widely-deployed.
We have also recently added a MATLAB/Simulink implementation of a Split-Band LPC-Based Speech Coder. We discuss some details of the project here.

Sample Course Project

After the labs are completed, the course also involves completion of a simple project which involves real-time simulation of a MATLAB/Simulink-based speech coding algorithm for coding speech [9].

Analog telephone systems have mostly been replaced by digital telephone systems. But with the advent of digital systems, the speech can be coded and has more flexibility, ease of regeneration and security than analog systems. But the disadvantage of digital systems is it requires larger bandwidth. The speech coding technology has gone through a number of phases starting with the development and deployment of PCM and ADPCM systems. The 64 kb/s Log-PCM and 32 kb/s ADPCM systems which have served the many early generations of digital systems well over the years have therefore been found to be inadequate in terms of spectrum efficiency when applied to the new, bandwidth limited, communication systems, e.g. satellite communications, digital mobile radio systems, and private networks. Hence in these systems signal coding and compression is vital. The latest trend is to reduce the bit rate from ADPCM's 32 kb/s to 2 kb/s LPC coders.

We now discuss an example of such a project which involves the encoder of the Split Band LPC based Vocoder in MATLAB Simulink.

A general sinusoidal analysis and synthesis concept was introduced by McAulay [10] when he developed the Sinusoidal Transform Coder (STC) [11] to demonstrate the applicability of the technique in low bit-rate speech coding. Sinusoidal model for the speech waveform is used to develop a new analysis/synthesis technique that is characterized by the amplitudes, frequencies, and phases of the component sine waves. One of the most recent harmonic coders operates in the LPC residual domain, i.e. Split Band LPC (SB-LPC) [13, 14].

Split Band LPC (SB-LPC):

The split-band linear predictive coding (SB-LPC) coder operating at 4 kb/s employs time-domain LPC filtering and uses a multi-band type of excitation signal. However the excitation signal of SB-LPC consists of only two bands, separated by a frequency marker, below which the spectrum is declared voiced and above which it is declared unvoiced. The estimation of the frequency marker of SB-LPC is different from the technique used in STC. The SB-LPC estimates a voicing decision for each harmonic band using a similar multi-band approach

SB-LPC replaces the binary excitation of the source-filter model with a more general mixed excitation, and filters the excitation signal using an LPC filter. The LPC residual has a simpler phase spectrum than the original speech. The Split band LPC coder splits the LPC excitation into frequency bands using a variable cut-off frequency

The Split band LPC coder will be studied and implemented in MATLAB Simulink as a course project. The coder has been tested for a small part of speech waveform. The waveform was voiced, unvoiced and mixed frames. The theory and the implementation of encoder, decoder and its sub-block is first discussed. The Simulation results are also described for each sub-block in the next sections.

Split Band LPC Encoder:

Split band encoder is depicted in the block diagram Fig. 2.1(a). It consists mainly of LPC encoder and the Sinusoidal encoder. From the block diagram, the input to the encoder is speech signal. The speech signal is sampled at 8kHz and is divided into 20ms frames. Each frame consists of 160 samples. The speech signal is passed through a high pass filter for DC rejection and is high frequency pre-emphasized. LPC parameters are determined using10th order Durbin's algorithm which are then quantized in the LSF domain. The quantized LPC parameters are then sent to LPC Inverse filter to find the LPC residual. The LPC residual is transformed to frequency domain using a 512-point FFT and this is used for determination of the excitation harmonic amplitudes.

The Pre-emphasized speech frame is transformed into frequency domain using Short-Time FFT. The Pitch detector block determines the pitch of the speech frame. Pitch block is discussed in detail in the next section. Voicing decision block determines the frequency marker up to which the frame is voiced. It uses the pitch information from the pitch block. Finally the harmonic amplitudes are determined from the excitation spectrum, using the weighted spectral matching and the LSF, pitch, voicing and the excitation parameters are finally quantized and transmitted to the decoder.

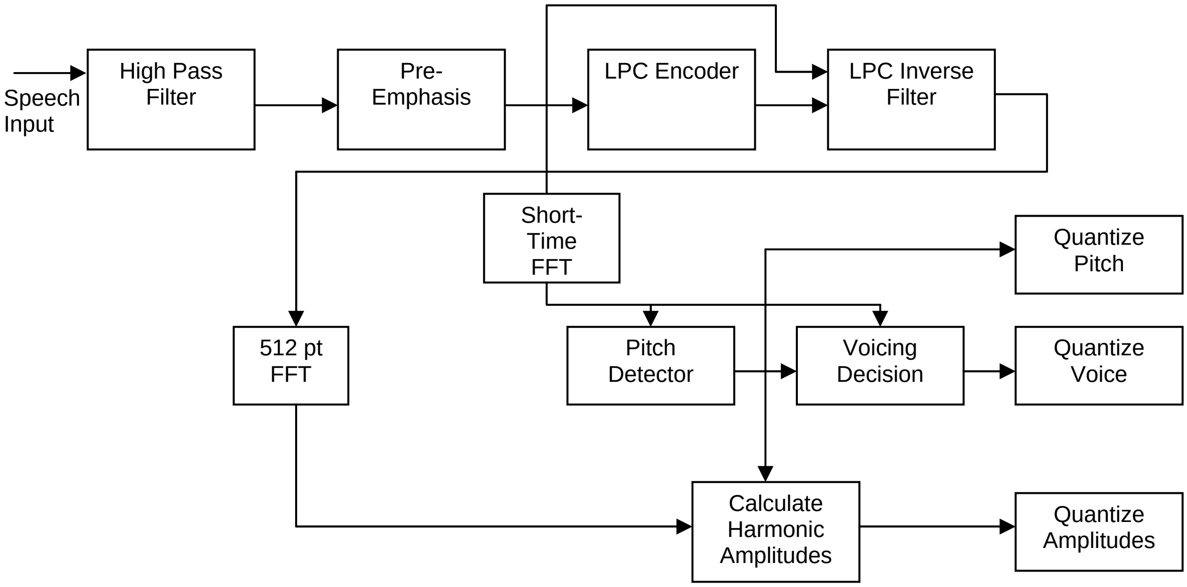


Fig 2.1(a): Block Diagram of Split-Band LPC Encoder.

The blocks are implemented in MATLAB Simulink. The Simulink Model is as shown in Fig.2.1(b)

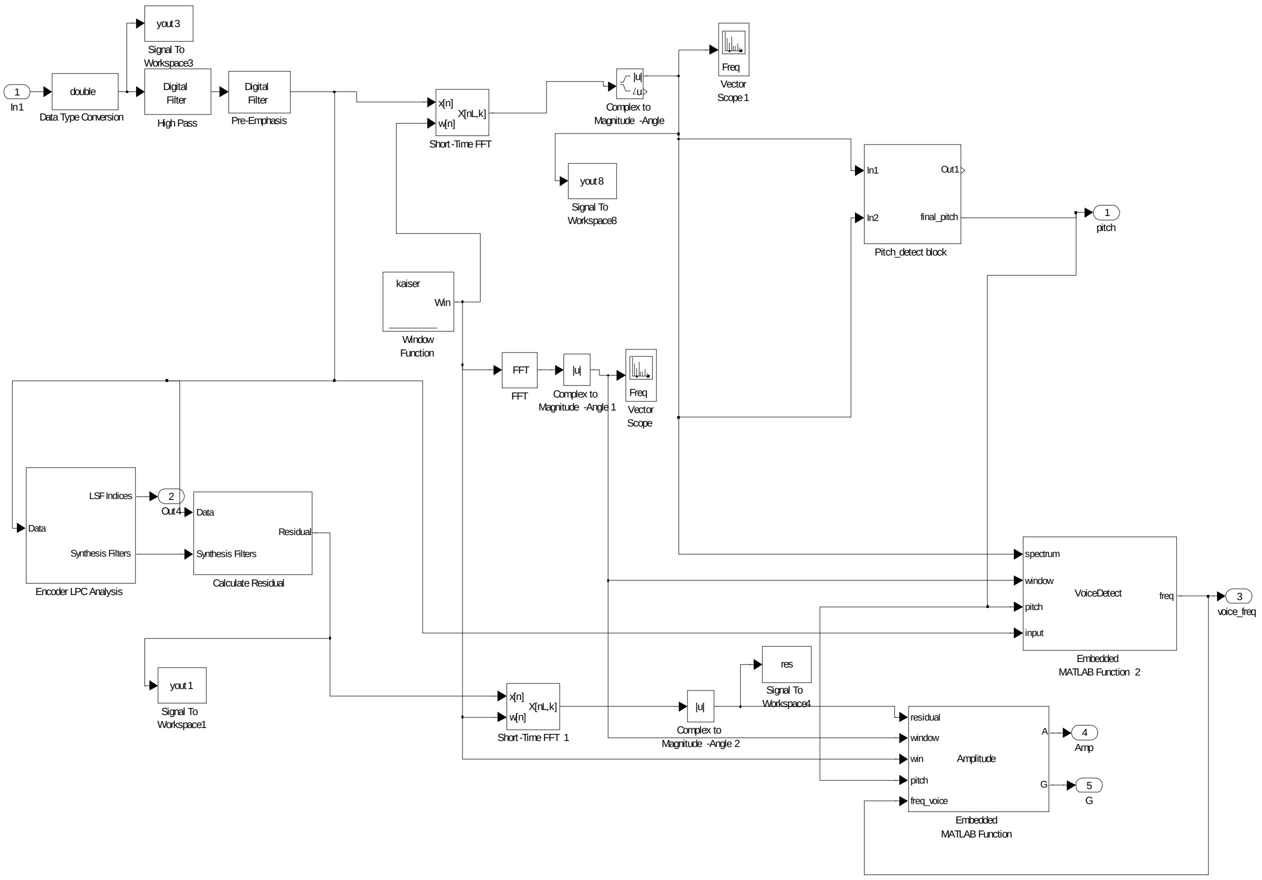


Fig 2.1(b): MATLAB Simulink model of SB-LPC Encoder

All the blocks, Pitch, Harmonic Amplitude, LPC Encoder and Voicing blocks are discussed in detail in the next section.

Encoder: Short Time FFT:

In the Short Time FFT (STFT), the data to be transformed is broken up into chunks or frames (which usually overlap each other). Each frame is Fourier transformed, and the complex result is added to a matrix, which records magnitude and phase for each point in time and frequency. This can be expressed as

$$\mathbf{STFT}\left\{x[n]\right\} \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \qquad (1)$$

Where, signal *x*[*n*] is speech frame and *w*[*n*] is the window. In this case, *m* is discrete and ω is continuous, but in most typical applications the STFT is performed on a computer using the Fast Fourier Transform, so both variables are discrete and quantized. The window function *w*[*n*] used is 221 point Kaiser Window which is generated by using

Generate Window block from MATLAB Simulink. The output of the window is as shown-

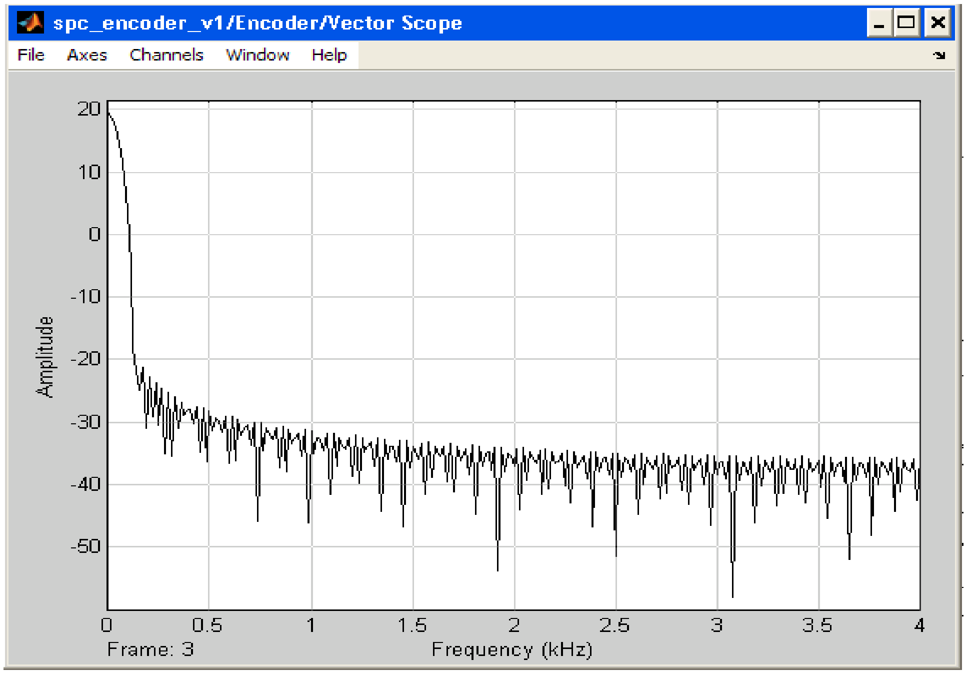


Fig 2.2(a) Frequency Spectrum of Kaiser Window

The Short Time FFT output of a frame from the MATLAB Simulink is shown below-

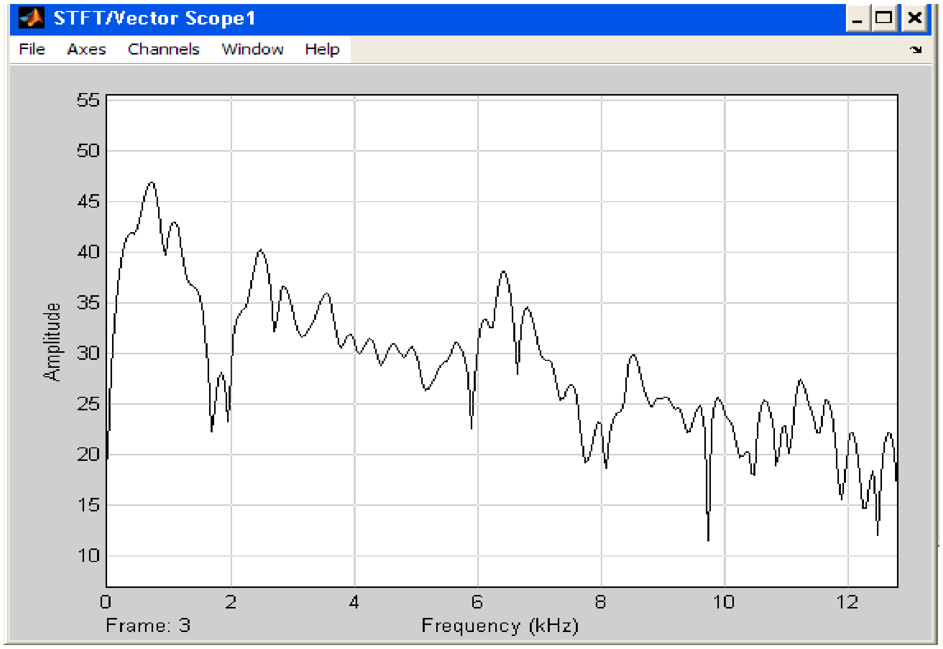


Fig 2.2(b) Speech Spectrum Output of the Short Time FFT

Encoder: Pitch Detector:

Pitch Analysis is performed in the spectral domain using the algorithm described by McAulay [10] which determines the pitch period to half sample accuracy. This algorithm estimates the pitch of a speech waveform that fits the harmonic set of sine waves to the input data using a mean squared error. The frame of the input speech waveform is analyzed in terms of its sinusoidal components. The speech measured data, s(n) can be represented as-

$$s(n)=\sum_{l=1}^{L}A_l\exp[j(nw_l+\theta_l)] \qquad (2)$$

Where $\{A_l,w_l,\theta_l\}_{l=1}^{L}$ represent the amplitudes, frequencies and phases of the L measured sine waves. The estimated sinusoidal waveform for which all of the frequencies are harmonic is given by-

$$\hat{s}(n;w_0,\phi)=\sum_{k=1}^{K(w_0)}\bar{A}(kw_0)\exp[j(nkw_0+\phi_k)] \qquad (3)$$

Where $w_0=2\pi f_0/f_s$ is the fundamental frequency, $K(w_0)$ is the number of harmonics in the speech bandwidth, $\bar{A}(kw_0)$ is the vocal tract envelope, ϕ is the phases of the harmonics. It is desired to estimate w_0 such that $\hat{s}(n)$ is as close as possible to s(n). This is achieved by calculating the minimum of minimum mean square (MSE),

$$\varepsilon(w_0,\phi)=\frac{1}{N+1}\sum_{n=-N/2}^{n=N/2}|\,s(n)-\hat{s}(n;w_0,\phi)\,|^2 \qquad (4)$$

After approximation, the above equation reduce to,

$$\varepsilon(w_0)=P_s-\rho(w_0) \qquad (5)$$

Where P_s is the power of the measured signal given by,

$$P_s=\sum_{l=1}^{L}A_l^2$$

Since the first term of equation (5) is a constant, the minimum mean squared error (MMSE) is obtained by maximizing $\rho(w_0)$ over w_0

If w_0 is the true pitch, then there would be at most one measured sine wave in each harmonic lobe tuned to w_0. Each lobe are determined by pitch adaptive sinc function, since each lobe spans one harmonic interval defined by the set-

$$L(kw_0)=\{w:kw_0-\frac{w_0}{2}\le w<kw_0+\frac{w_0}{2}\} \qquad (6)$$

The MSE pitch estimation $\rho(w_0)$ criterion becomes-

$$\rho(w_0) = \sum_{k=1}^{K(w_0)} \bar{A}(kw_0)\{\max_{w_l \in L(kw_0)} [A_l D(w_l - kw_0)] - \frac{1}{2}\bar{A}(kw_0)\} \qquad (7)$$

Where $D(w_l - kw_0) = \dfrac{\sin[2\pi(\frac{w-kw_0}{w_0})]}{2\pi(\frac{w-kw_0}{w_0})}\ \mathit{forall}\ |w - kw_0| \leq \dfrac{w_0}{2}$ (8)

The above equation is calculated for all values of w_0 and the pitch is estimated for the maximum value of $\rho(w_0)$. The enhanced MSE criterion insures that speech of low pitch will less likely be estimated as a high pitch.

Sine wave envelope is estimated using SEEVOC method which is discussed in the next section.

Sine-Wave Amplitude Envelope Estimation:

The MSE criterion $\rho(w_0)$ can lead to unambiguous estimates of the pitch if the sine-wave amplitudes are known. One of the methods used is Spectral Envelope Estimation Vocoder (SEEVOC). In this method we assume the value of average pitch $\bar{w}_0$. And the first step is to find the largest amplitude in the frequency range$[\frac{\bar{w}_0}{2}, \frac{3\bar{w}_0}{2}]$. After finding the amplitude and frequency of that peak, denoted by (A_1, w_1), then the interval $[w_1 + \frac{\bar{w}_0}{2}, w_1 + \frac{3\bar{w}_0}{2}]$ is searched for its largest peak, (A_2, w_2). The process is continued throughout the speech band. The envelope is formed by linearly interpolating between the successive log amplitudes using the peak values determined by the above search procedure.

SIMULINK Implementation:

The input to the Pitch detector is the 512 point Short-Time FFT output of the speech signal as shown below. The sine wave amplitudes and frequencies were determined over 4000Hz bandwidth. The block implemented is shown in Fig (2.3 (a)).

Peak Finder block is from the Simulink Library which selects 100 Peak Values and the corresponding frequencies from the 512-point FFT speech signal. All the peak values and the frequencies are sent to the MSE pitch extractor. This block uses the peak values of the speech signal and the speech envelope estimated from the amplitude block to calculate the final pitch for the frame.

In the pitch extractor block, it calculates the MSE criterion $\rho(w_0)$ (Equation 7) for pitch values from 38 Hz to 400 Hz. It should be noted that the peak at the correct pitch is the largest. The speech envelope is calculated in the amplitude block. Here we assume the average pitch $\bar{w}_0$ to be 200Hz.

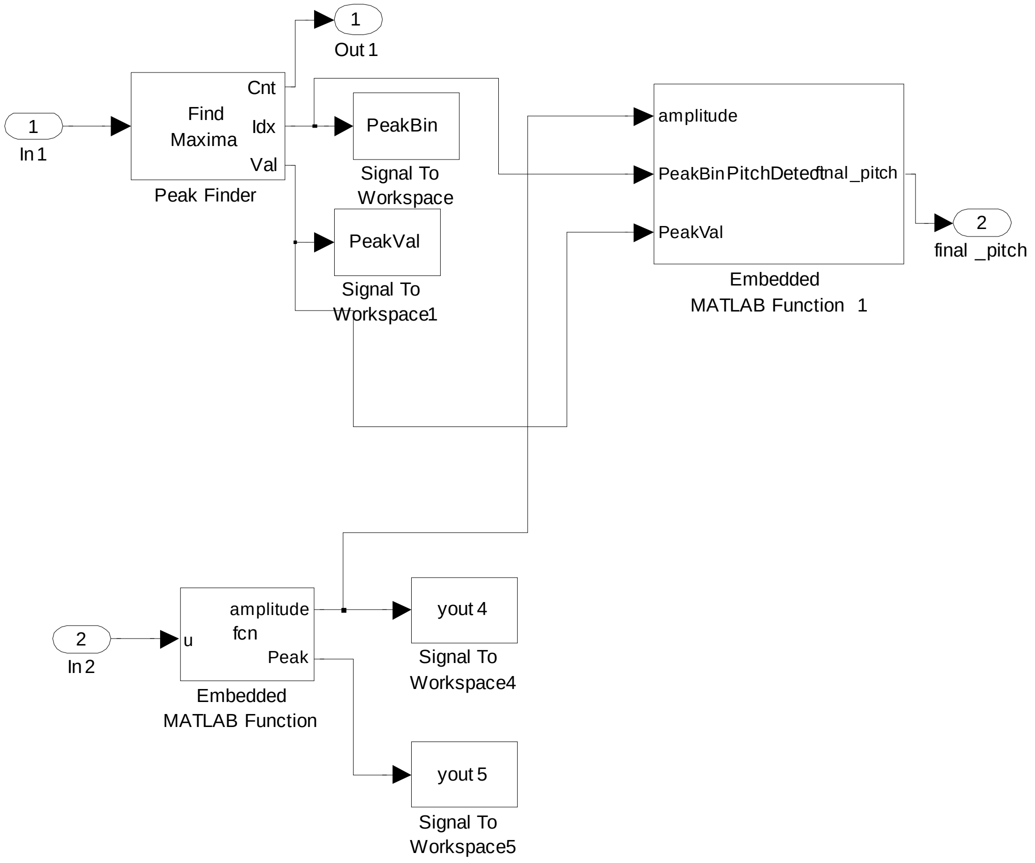
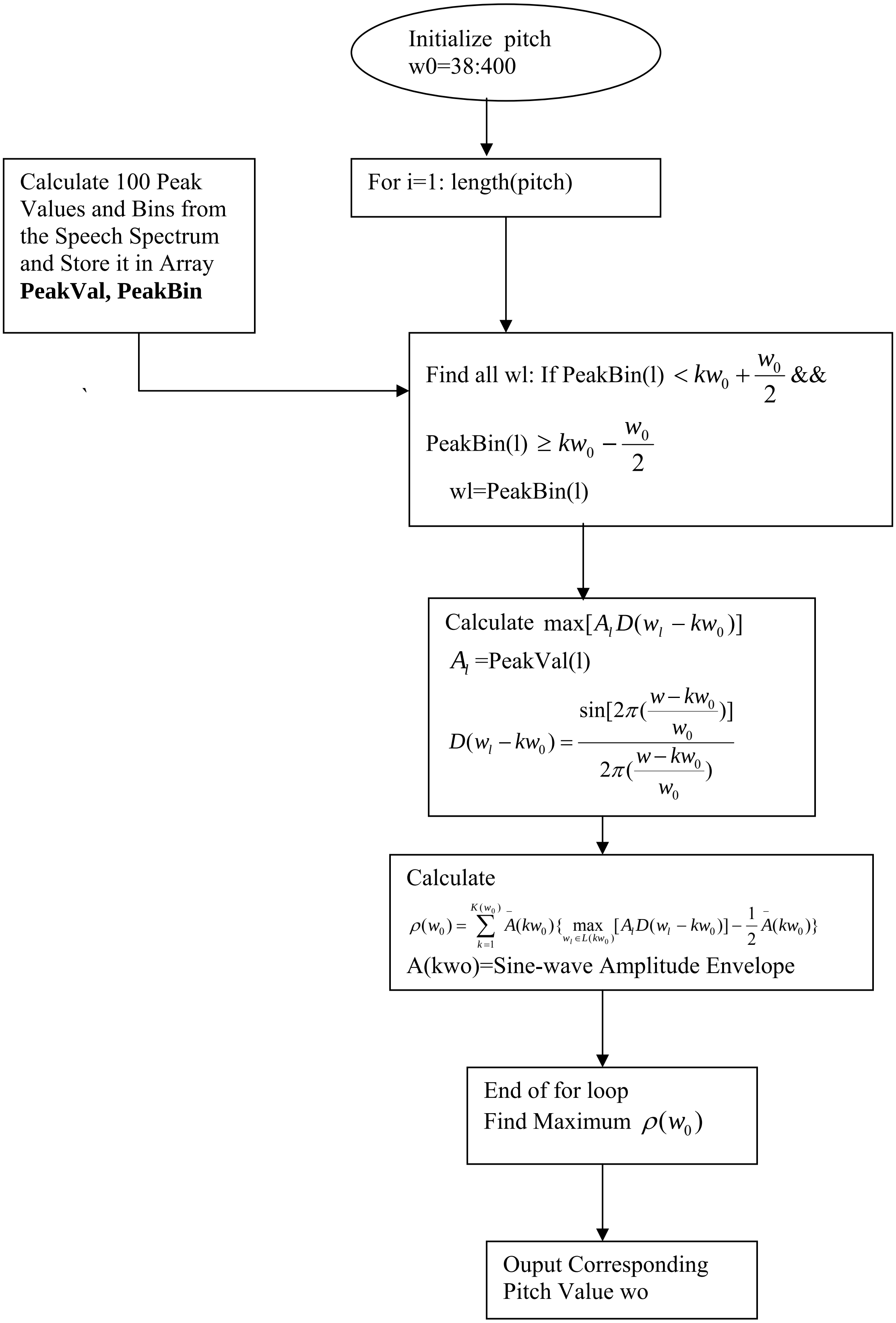


Fig 2.3 (a) Pitch detector block SB_LPC encoder in MATLAB Simulink

Fig 2.3 (b) Flow chart of the Pitch Detector Block


Initialize pitch
w0=38:400
For i=1: length(pitch)
Calculate 100 Peak Values and Bins from the Speech Spectrum and Store it in Array **PeakVal, PeakBin**
Find all wl: If PeakBin(l) $< kw_0 + \frac{w_0}{2}$ &&
PeakBin(l) $\geq kw_0 - \frac{w_0}{2}$
wl=PeakBin(l)
Calculate $\max[A_l D(w_l - kw_0)]$
A_l =PeakVal(l)
$D(w_l - kw_0) = \frac{\sin[2\pi(\frac{w-kw_0}{w_0})]}{2\pi(\frac{w-kw_0}{w_0})}$
Calculate
$\rho(w_0) = \sum_{k=1}^{K(w_0)} \bar{A}(kw_0)\{\max_{w_l \in L(kw_0)}[A_l D(w_l - kw_0)] - \frac{1}{2}\bar{A}(kw_0)\}$
A(kwo)=Sine-wave Amplitude Envelope
End of for loop
Find Maximum $\rho(w_0)$
Ouput Corresponding Pitch Value wo

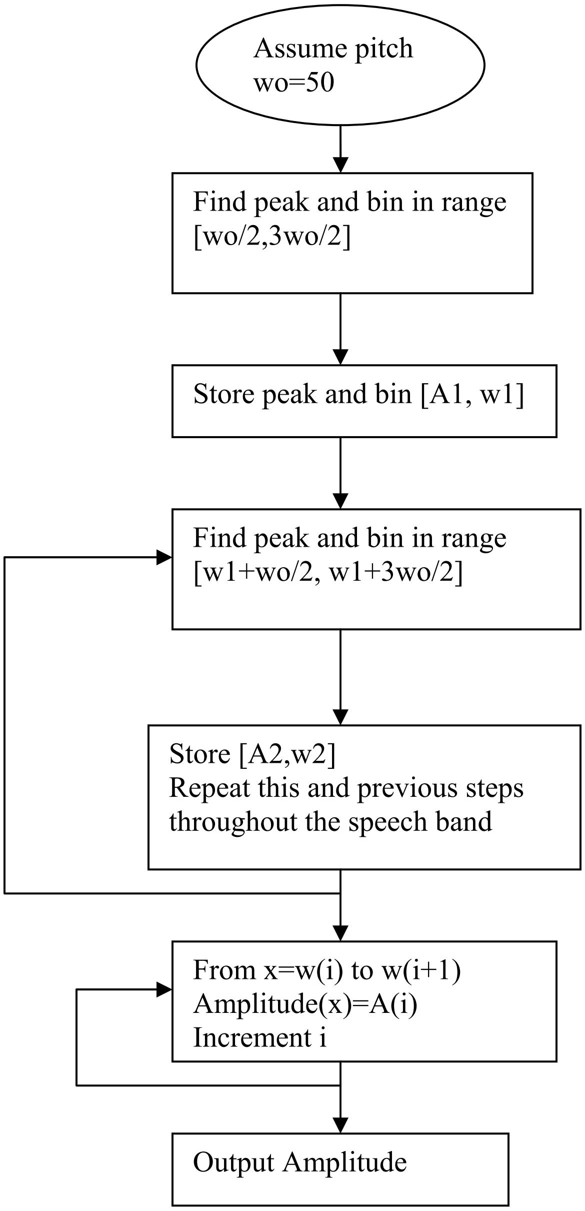

Assume pitch
wo=50
Find peak and bin in range
[wo/2,3wo/2]
Store peak and bin [A1, w1]
Find peak and bin in range
[w1+wo/2, w1+3wo/2]
Store [A2,w2]
Repeat this and previous steps
throughout the speech band
From x=w(i) to w(i+1)
Amplitude(x)=A(i)
Increment i
Output Amplitude

Fig 2.3 (c) The Flow chart for the Amplitude Sine wave envelope

The input speech waveform used to test the Pitch Detector block was a voiced signal with frame length of 3 and the number of samples in each frame is 160 samples. From the diagram in Fig 2.3 (d), we can see that it is periodic with period of 100 Hz.

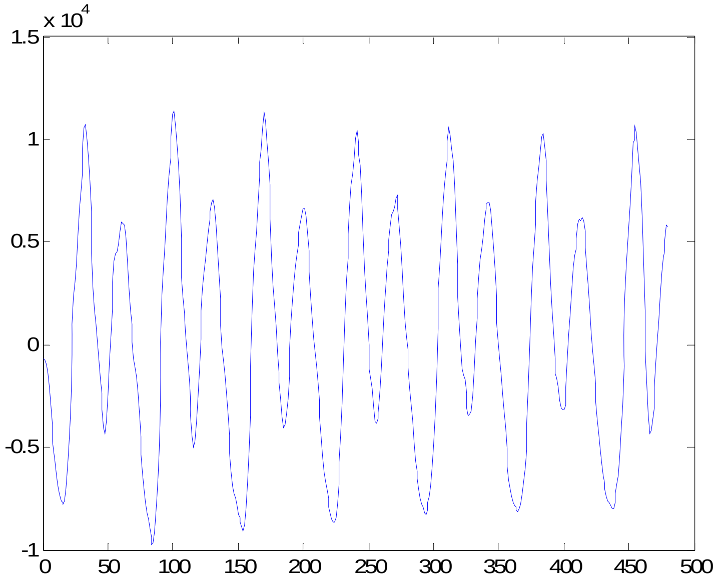


Fig 2.3 (d) Voiced Speech Input Signal

The measured amplitudes and frequencies are shown along with the piecewise constant SEEVOC envelope as in Fig. 2.3(e). The signal in red is the SEEVOC envelope and the green signal is the speech spectrum. We can see that the envelope follows the speech spectrum with the pitch frequency harmonics.

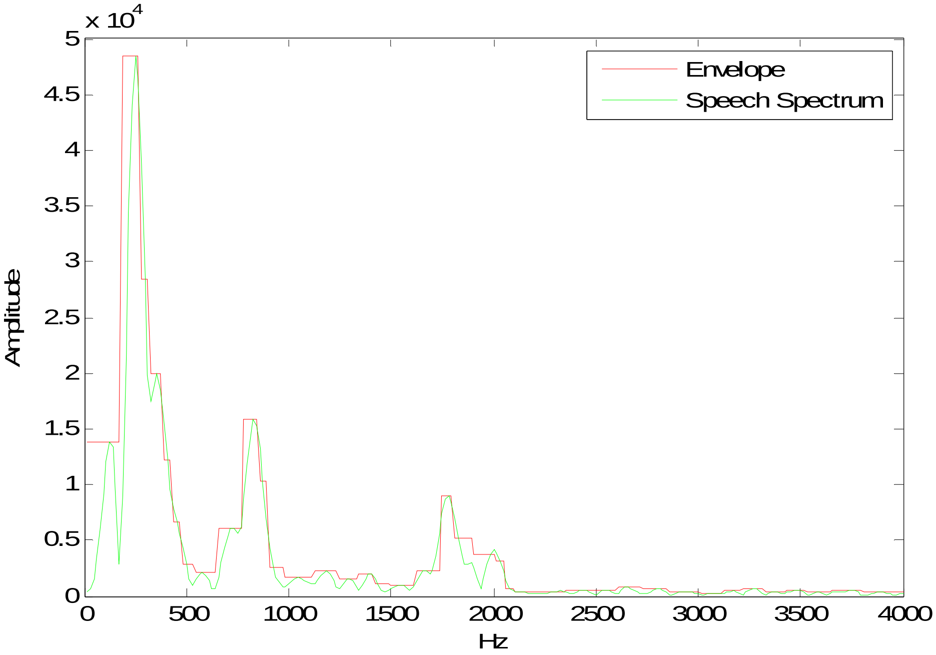


Fig 2.3 (e) Sine Wave Amplitude Envelope Estimation

The overall MSE criterion $\rho(w_0)$ is plotted over the pitch range from 38Hz to 400Hz as in Fig. 2.3(f). The maximum value from the figure was found to be around 110Hz which conforms to the input signal.

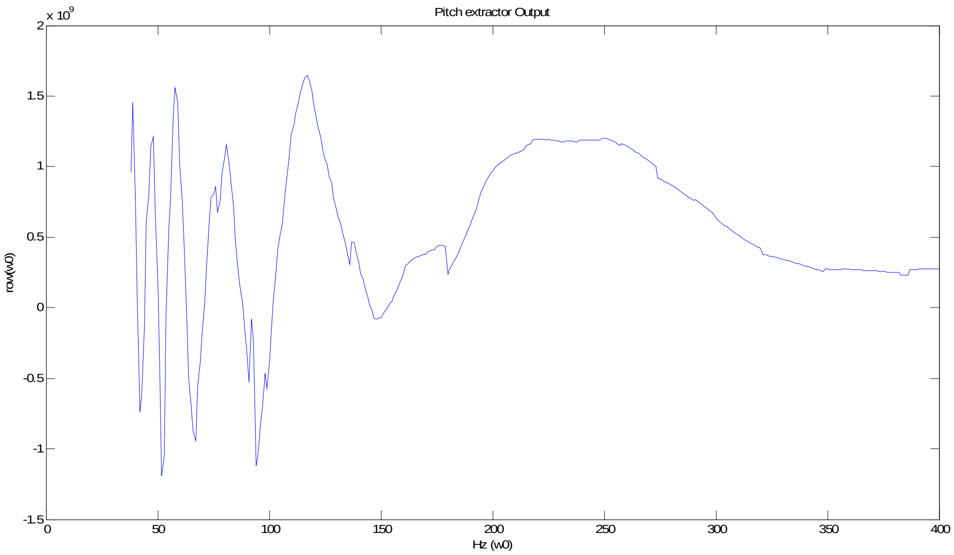


Fig 2.3 (f) Overall MSE criterion versus frequency (Hz)

Conclusions

The area of Multimedia Systems has become an important application area of Digital Signal Processing (DSP) in the Computer and Electrical Engineering curriculum
This include areas of speech, audio and video data processing.
We are developing a 3-part course sequence that will help teach undergraduates multimedia systems. The first part of the 3-part course sequence is about speech. The second and third parts will focus on audio and video.

In this paper, we provided details of the course and labs we designed for undergraduates which focuses on understanding how to process speech. We provide examples of the curriculum, what is covered and how we cover it. We also provide examples of an end-of-class projects using a Split-Band LPC speech coder.

In the future, we plan to present work on other two courses under development and learning outcomes, assessment tests, etc. An overview of applications of signal processing in voice-over-IP networks is discussed in [15]. This can also be used as example.

References

[1] T. Ogunfunmi, Pedagogy of a course in Speech Coding and Voice-over-IP, *Proceedings of the ASEE Annual Conference,* June 2008.
[2] T. Ogunfunmi and M.J. Narasimha, Principles of Speech Coding. Santa Clara University, Santa Clara, CA, 1997-2007.

[3] American Board for Engineering and Technology (ABET) Engineering Accreditation Commission, *Criteria for Accrediting Engineering Programs*, Baltimore, MD, 2003. (URL: http://www.abet.org)

[4] Chu, Wai C., Speech Coding Algorithms: Foundation and Evolution of Standardized Coders, Wiley Publishers, 2003.

[5] Kondoz, A.M., Digital Speech: Coding for low bit rate communication systems, Wiley Publishers, (Second edition), 2004.

[6] Gibson, Jerry, Speech Coding Methods, Standards and Applications, *IEEE Circuits and Systems Magazine,* pp. 30-49, Fourth Quarter 2005.

[7] ITU-T Recommendation G.711 Standard. (http://www.itu.int).

[8] ITU-T Recommendation G.729 (03/96), Coding of Speech at 8 kbit/s using conjugate structure algebraic-code-excited linear-prediction (CS-ACELP).

[9] Ramya Nagaraj, Study and Implementation of Split-Band LPC Based Vocoder in MATLAB/Simulink, MS Thesis, Santa Clara University, June 2010.

[10] I. Atkinson, S. Yeldener, and A. Kondoz (1997) 'High quality split-band LPC vocoder operating at low bit rates', in *Proc. of Int. Conf. on Acoust., Speech and Signal Processing*, pp. 1559–62. May 1997. Munich

[11] Split Band LPC based Adaptive Multi-Rate GSM Candidate *S.Villette, M.Stefanovic, A.Kondoz* Centre for Communication Systems Research University of Surrey, Guildford GU2 5XH, Surrey, United Kingdom.

[12] D. W Griffin., J. **S.** Lim, "Multi-Band Excitation Vocoder", IEEE Trans. ASSP, Vol. 36, No.8, pp 1223-1235, Aug. 1988.

[13] R. J. McAulay, T. F. Quateri, "Pitch Estimation and Voicing Decision Based upon **a** Sinusoidal Speech Model", Proc. ICASSP, 1990.

[14] R. J. McAulay and T. F. Quatieri (1986) 'Speech analysis/synthesis based on a sinusoidal representation', in *IEEE Trans. on Acoust., Speech and Signal Processing*, 34(4):744–54

[15] T. Ogunfunmi and M.J. Narasimha, Speech over VoIP Networks: Advanced Signal Processing and System Implementation, accepted for publication, IEEE Circuits and Systems, Feb. 2012.