IJEEE
ISSN 2321 - 2055

SPEECH PROCESSING SYSTEM IN VLSI DESIGN USING LD ALGORITHM

A.LYDIA[1] , G.JAYANTHI[2]

[1]P.G Scholar, VLSI Design,Parisutham Institute of Technology and Science, (India)
[2]Assistant Professor Electronics and Communication Engineering,
Parisutham Institute of Technology and Science, (India)

ABSTRACT

Speech coding has been major issue within the space of digital speech process. It unfeasible to access unlimited bandwidth of a channel when we tend to send a signal across it that results in code and compress speech signals. Speech compression is needed within the areas of long distance communication, high-quality speech storage, and message coding. For instance in digital cellular technology several users ought to share a similar frequency bandwidth. Utilizing speech compression makes it attainable for a lot of users to share the accessible system. Another example wherever speech compression is required in digital voice storage for a hard and fast quantity of accessible memory compression makes it potential to store longer messages. Speech coding may be a lossy sort of coding, which suggests that the signaling doesn't specifically sound just like the input. Our work consists of truncating a recorded voice signal, framing it, passing it through a window function and estimation of Levinson Durbin algorithmic program.By increasing the PSNR we are able to cut back the error rate and conjointly it'll improve the standard of the system.

Keywords: Levinson Durbin, Speech compression, PSNR

I. INTRODUCTION

Data compression is a technique during which data content of the input to system is compressed so original signal is obtained as output and unwanted or unsought signals are removed. so once speech signals are utilized in the shape of data it is termed as SPEECH COMPRESSION. Speech is a actual basic method for humans to convey information to one another.

Speech includes a little bandwidth of 4 khz. Speech compression involves coding of real-time audio signals at all-time low attainable bit rates. The compression of speech signals has several sensible applications. it is utilized in digital cellular technology wherever several users share identical frequency bandwidth. Compression permits additional users to share the system at a selected time. It can even be used for digital voice storage that are used for responsive machines and pre-recorded phonephone calls that are used for purpose of providing any reasonable information to user or advertising. For a given memory size, compression permits longer messages to be stored than otherwise.

Compression is employed almost all over. All the sounds that is from web are compressed, most modems use compression, and a number of other file systems automatically compress files once stored, and therefore the rest of us make out by hand. we have a tendency to should distinguish between "lossless algorithms", which might reconstruct the original message specifically from the compressed message, and "lossy algorithms", which might solely reconstruct an approximation of the original message. lossless algorithms are generally used for text, and lossy for images and sound. clearly such an interface would yield nice advantages. attempts are created to develop vocally interactive computers to appreciate voice/speech recognition. during this case a system will acknowledge text and provides out a speech output.

II. LITERATURE SURVEY

[1] In acoustic modeling, speaker adaptive training (SAT) has been a long-standing technique for the normal Gaussian mixture models (GMMs). Acoustic models trained with SAT become freelance of coaching speakers and generalize better to unseen testing speakers. This paper ports the concept of SAT to deep neural networks (DNNs), and proposes a framework to perform feature-space sat for DNNs. victimization i-vectors as speaker representations; our framework learns an adaptation neural network to derive speaker-normalized options. Speaker adaptive models are obtained by fine-tuning DNNs in such a feature area. These frameworks are often applied to numerous feature types and network structures, posing a very general sat resolution.

[2] Speech emotion recognition, which is defined as extracting the emotional states of a speaker from his or her speech, is attracting more attention. It is believed that speech emotion recognition can improve the performance of speech recognition systems and is thus very helpful for criminal investigation, intelligent assistance surveillance and detection of potentially hazardous events and health care systems. Speech emotion recognition is particularly useful in man-machine interaction. Speech emotion recognition is found in our study that the first- and second-order differences of harmony options additionally play a vital role in speech emotion recognition. Therefore, we have a tendency to propose a replacement Fourier parameter model exploitation the perceptual content of voice quality and therefore the first- and second-order variations for speaker-independent speech emotion recognition. Experimental results show that the proposed Fourier parameter (FP) features are effective in distinguishing numerous emotional states in speech signals.

[3] As a vital approach of human emotional behaviour understanding, speech emotion recognition (SER) has attracted a great deal of attention in humanistic signal process. Accuracy in SER heavily depends on finding sensible affect-related, discriminative features. During this paper, we have a tendency to propose to find out affect- salient features for SER exploitation convolutional neural networks (CNN). The training of CNN involves 2 stages. Within the initial stage, unlabeled samples are used to learn native invariant options using a variant of thin auto-encoder with reconstruction penalization. in the second step, LIF is employed because the input to a feature extractor, salient discriminative feature analysis, to learn affect-salient, discriminative features employing a novel objective function that encourages feature salience, orthogonality, and discrimination for SER. The second research issue is to determine the most suitable type of features for SER.

[4] We recently presented an efficient approach for training a Pairwise Support Vector Machine (PSVM) with an appropriate kernel for i–vector pairs for a quite massive speaker recognition task. Instead of estimating an SVM model per speaker, according to the “one versus all” discriminative paradigm, the PSVM approach classifies an attempt, consisting of a combine of i–vectors, as belonging or not to a similar speaker category. Training a PSVM with great amount of data, however, is a memory and machine overpriced task, as a result of the number of training pairs grows quadratically with the quantity of training i–vectors. This paper demonstrates

that a very little subset of the training pairs is necessary to train the initial PSVM model, and proposes two approaches that permit discarding most of the training pairs that don't seem to be essential, while not harming the accuracy of the model.

[5] Humans will extract speech signals that they have to understand from a mixture of background, intrusive sound sources, and reverberation for effective communication. Voice Activity Detection (VAD) and Sound source Localization (SSL) are the key signal process elements that humans perform by processing sound signals received at each ears, sometimes with the help of visual cues by locating and perceptive the lip movements of the speaker. The design and implementation of robust VAD and SSL algorithms in practical acoustic environments are still difficult issues, notably once multiple simultaneous speakers exist within the same audiovisual scene. In this work we tend to propose a multimodal approach that uses Support Vector Machines (SVMs) and Hidden Markov Models (HMMs) for assessing the video and audio modalities through an RGB camera and a microphone array.

III. EXISTING SYSTEM

Support vector machine (SVM)-based speaker verification system. The desire chip includes a speaker feature extraction (SFE) module, an SVM module, and a decision module. The SFE module performs autocorrelation analysis, linear predictive coefficient extraction, and LPC-to-cepstrum conversion. The SVM module includes a Gaussian kernel unit and a scaling unit. an intense CORDIC design is proposed to calculate the exponential value. moreover because the Gaussian kernel unit, a scaling unit is additionally developed to be used within the SVM module. The scaling unit is employed to perform scaling multiplications and therefore the remaining operations of SVM decision value analysis. Finally, the decision module accumulates the frame scores that are generated by all of the check frames, and so compare it with a threshold to visualize if the check utterance is spoken by the claimed speaker. This designed chip is characterised by its high speed and its ability to handle an oversized range of support vectors within the SVM.

IV. PROPOSED SYSTEM

The objective is to minimize transmission prices or offer cost efficient storage. High rate of compression with high PSNR got with so closed speech signal to the original one, measured using correlation factor which provides values almost one. Specifically to improve the standard of speech by means that of accelerating the PSNR by decreasing the MSE value. Peak signal-to-noise ratio, typically abbreviated PSNR, may be a term for the ratio between the maximum attainable power of a signal and therefore the power of corrupting noise that affects the fidelity of its illustration. as a result of several signals have a really wide dynamic range, PSNR is typically expressed in terms of the logarithmic dB scale. To save a memory space, a voice is digitized and compressed using LD algorithmic rule. Data compression may be a technique in which data content of the input to system is compressed in order that original signal is obtained as output and unwanted or unsought signals are removed.

4.1 System Overview

The proposed speaker verification chip consists mainly of a SFE module, an LD module, a decision module, and a control module as shown in fig 1.The advantage of the core-based design for SFE and LD is its flexibility that allows the designer to develop a new system in short time using suitable IP cores. For the SFE module, since the adjacent frames are overlapping, an intelligent architecture is used to perform the autocorrelation analysis without the use of a huge buffer. This architecture allows the input buffer for the SFE module to contain only two registers instead of a huge buffer. The detail of the SFE module can be found in.

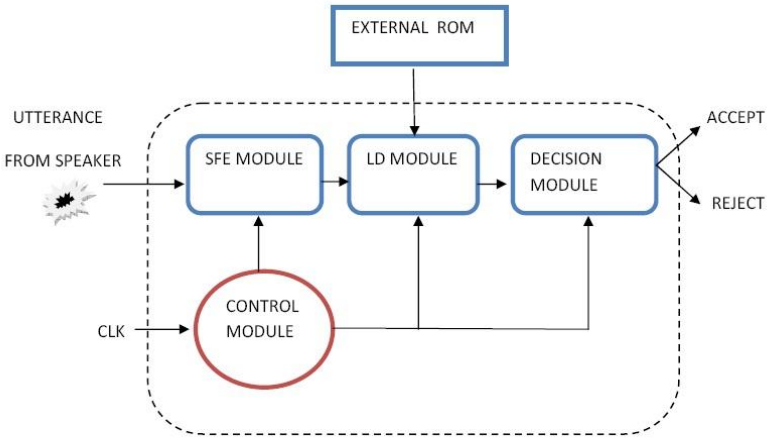

Fig 1: Proposed system block diagram

The decision module can, thereby compute the overall score by summing all of the frame scores. This obtained overall score is compared with a threshold (higher threshold of e.g. −5 dB) and lower threshold (e.g. -60 dB) to determine whether the test utterance is compressed at a high rate.

4.2 Overall Block Diagram

The figure 2 shows the overall block diagram, let us discuss each module in detail based on the transmission conditions.

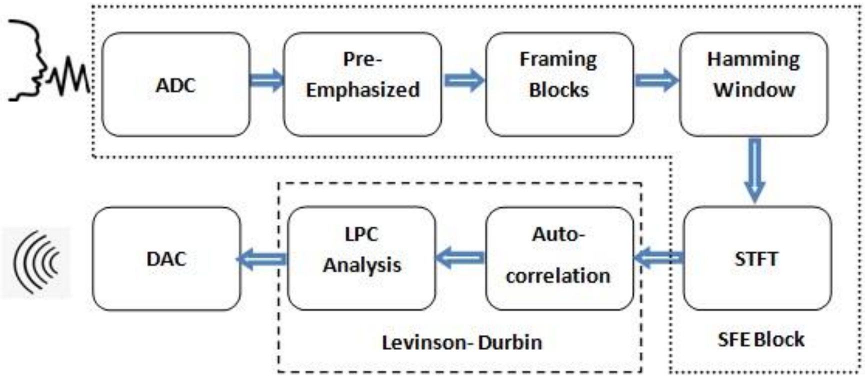

Fig 2: Overall block diagram

4.3 Functional Module For The Sfe Block

The functional module for SFE block is shown in fig 3 consists of ADC, Pre-emphasize filter, Frame block, hamming window and STFT. The analog-digital converter (ADC) interprets this analog wave into digital information that the system will perceive. To do this, it samples, or digitizes, the sound by taking precise measurements of the wave at frequent intervals.

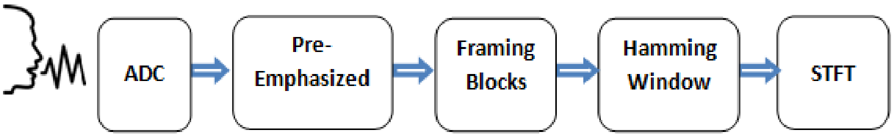

Fig 3: SFE Module

Pre-emphasis is a simple straightforward signal process methodology that will increase the amplitude of high frequency bands and reduces the amplitudes of lower bands. The pre-emphasized speech samples are divided into 30ms window frames. Consequent step is to window every individual frame, here hamming window is employed to reduce the signal discontinuities at the start and finish of every frame. The idea applied here is to attenuate the spectral distortion by victimization the window to taper the signal to zero at the start and finish of every frame. If we have a tendency to outline the window as w (n), 0 ≤ n ≤ N -1, wherever N is that the frame length, then the results of windowing is that the signal.

$$y(n) = x(n)w(n), 0 \le n \le N - 1 \qquad (1)$$

We have used the Hamming window in our project, which has the form:

$$0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \; 0 \le n \le N - 1 \qquad (2)$$

Where,

x (n) = Input signal

w (n) = Window function

N = Frame length

For STFT, we impose window of certain size onto the original signal, then we perform FFT on each window FFT, which converts each frame of N samples from the time domain into the frequency domain.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N} \;, k = 0,1,2, \dots, N-1 \qquad (3)$$

Where,

X_k = Frequency

N = Frame length

4.4 Levinson Durbin Algorithm

The formal method for converting from autocorrelation coefficients to an LPC parameter set is known as Durbin’s method. The Levinson–Durbin algorithm is a recursive order-update method for calculation of linear predictor coefficients. Fig 4 shows the LD algorithmic flow, which is solved by Yule walker equations.

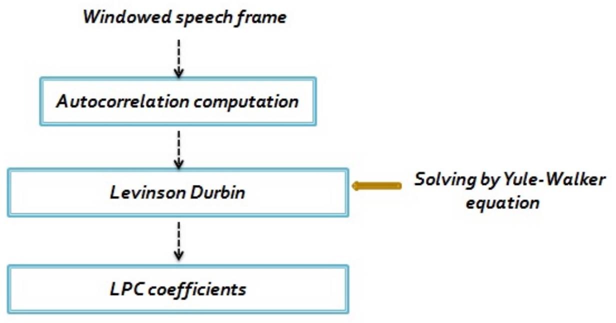

Fig 4: LD algorithmic flow

Autocorrelation: In autocorrelation methodology, a window is employed to divide the speech into. For every placement of window, sometimes from 10 to 30 ms apart, the speech signal is windowed to make one analysis frame of the signal. The feature of Autocorrelation is primarily completely different for voiced and unvoiced segments of speech. Therefore information from the autocorrelation sequence will be used for discriminating voiced and unvoiced segments. Consequently, the autocorrelation of voiced speech ought to offer sturdy peak at the periodic worth and no such peak just in case of unvoiced speech. Therefore, the autocorrelation of speech has become a customary approach for enhancing pitch.

$$R\ (t1,t2) = E\ \{x\ (t1)\ x\ (t2)\} \quad (4)$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1 x_2 f(x_{1,} x_2; t_1 t_2) dx_1 dx_2 \quad (5)$$

$$r_{x\,(h)} = \sum_{n=0}^{N-h-1} x^*(n) x(n+h) \quad h = 0,1, \ldots N-1 \quad (6)$$

where,

x(t) is the expected value of pitch, h is the lag between samples,* denotes complex conjugate, N is the input length and rx(h) indicates the theoretical autocorrelation

Zero Crossing Rate (ZCR): The input speech signal can be viewed in blocks of 10-30 ms for computing ZCR. For each block of the speech signal, the ZCR is computed using the short term ZCR relation. The ZCR value is highest in unvoiced region and lowest in voiced region. In case of silence region the value lies in between of voiced and unvoiced cases. Hence magnitude sum function (msf) is that adding two vectors (magnitude).

$$|a| = |b|\ \ |c| \quad (7)$$

LD: The Levinson Durbin block solves the nth order system of linear equations, hence R is a Hermitian, positive definite, Toeplitz matrix.

$$Ra = r \quad (8)$$

$$\begin{bmatrix} r(0) & r(1) & \cdots & r(M-2) & r(M-1) \\ r(1) & r(0) & \cdots & r(M-3) & r(M-2) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ r(M-2) & r(M-3) & \cdots & r(0) & r(1) \\ r(M-1) & r(M-2) & \cdots & r(1) & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{M-1} \\ a_M \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(M-1) \\ r(M) \end{bmatrix} \quad (9)$$

For calculating the single coefficients where P = Predictor of order. This is known as the Levinson-recursion.

$$a_i^{(P)} = a_i^{(P-1)} - a_p^{(P)} a_{p-i}^{(P-1)}\ for\ i = 1, \ldots.P-1 \quad (10)$$

Energy of the predictor error signal, where Eerror is the prediction error,

$$E_{error}^{(P)} = E_{error}^{(P-1)} - (1 - a_p^{(P)2}) \quad (11)$$

Initial Values are,

$$a_0^{(0)} = 1 \qquad (11)$$

$$E_{error}^{(0)} = r_0 \qquad (12)$$

One application of the Levinson Durbin formulation enforced by this block is in the Yule Walker PSD using Yule-Walker technique: this can be a special case of the smallest amount squares method wherever the whole error residuals are used. These ends up in diminished performance compared to the covariance strategies, however is also expeditiously resolved using the Levinson formula. It is also called the autocorrelation technique

V. RESULTS AND DISCUSSION

Synthesized speech can be created by concatenating pieces of recorded speech that are stored. Systems differ in the size of the stored speech units. The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood clearly

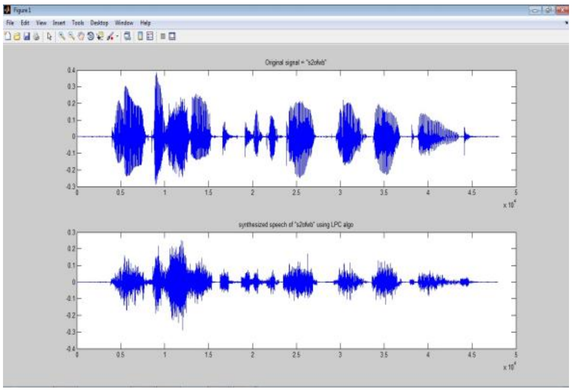

Fig 5: Synthesized speech

Table 1 shows the Compression Ratio, MSE, PSNR, TIME and Correlation Coefficient

FILE NAME	COMPRESSION RATIO	MSE	PSNR	TIME	CORRELATION COEFFICIENT
s2ofwb	0.75	1.8311e-04	69.1336	114.8813	0.9970
s1ofwb	0.05	3.6621e-04	56.3525	416.0365	0.9982
kdt_070	0.43	5.1880e-04	41.4524	302.0199	0.9952
kdt_003	1	2.1362e-04	61.2535	444.6526	0.9948
Average	**0.522**	**5.8311e-04**	**56.1336**	**319.8813**	**0.9970**

The parameter of table (1) shows that, as the MSE value reduces the PSNR value increases, hence the correlation coefficient value will be nearer to 1and thus compression values average is about 0.5 with respect to time.

VI. CONCLUSION

Due to the practical application on the suggested algorithm, The MSE and PSNR are used as fidelity criteria during the test phase. The following can be seen as conclusion for the suggested approach: High rate of compression can be got with medium size of speech file. Retrieved speech still with high quality low noise and so closed to the original one. Difficult for listener to recognize the effect of the difference between the original and retrieved speech. Efficient speech compression can be achieved with the help of linear predictive coding. The results obtained using LD algorithm these compression ratios are expressed in decibels, so that a ratio of 2:1 indicates that a signal exceeding the threshold by 2 dB will be attenuated down to 1 dB above the threshold, or a signal exceeding the threshold by 8 dB will be attenuated down to 4 dB above it. We have achieved 50% of data compression. And the PSNR rate increases with decrease in MSE. Future versions of the system are planned to consider two important aspects in speaker identification they are the inter-speaker distance and the intra-speaker variability with different emotions and finding the respective user using Gammatone Filtering and Cochleagram Coefficients in XILINX ISE software. Hence the results will be compared with the existing system, which indicate that the system can achieve significant performance

REFERENCES

[1] Arjona Ramírez M ,M. Minami, “Technology and standards for low-bit-rate vocoding methods,” in The Handbook of Computer Networks, H. Bidgoli, Ed., New York: Wiley, 2008, vol. 2, pp. 447–467

[2] BenZeghiba M. F, “Joint speech and speaker recognition,” M.S. thesis, Dept. Comput. Sci., Swiss Federal Inst. Technol. Lausanne (EPFL), Lausanne, Switzerland, 2005

[3] Bing Nan Li, Senior Member, IEEE, Kunxia Wang, Member, IEEE, Lian Li, Ning An, Senior Member, IEEE, Yanyong Zhang, Member, IEEE, “Speech Emotion Recognition Using Fourier Parameters,” IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 6, NO. 1, JANUARY-MARCH 2015

[4] Bowon Lee, Senior Member, IEEE, Claudio R. Jung, Senior Member, IEEE, Vicente P. Minotto, “Simultaneous-Speaker Voice Activity Detection and Localization Using Mid-Fusion of SVM and HMMs,” IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 16, NO. 4, JUNE 2014

[5] Campbell W. M, D. A. Reynolds and, D. E. Sturim, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Process. Lett., vol. 13, no. 5, pp. 308–311, May 2006

[6] Chen T. H and B. H. Juang, “The past, present, and future of speech processing,” IEEE Signal Process. Mag., vol. 15, no. 3, pp. 24–48, May 1998.

[7] Christodoulou C. C, T. G. Clarkson, D. Gorse, Y. Guan, D. A. Romano-Critchley, and J. G. Taylor, “Speaker identification for security systems using reinforcement-trained pRAM neural network architectures,” IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 31, no. 1, pp. 65–76, Feb. 2001

[8] Bryan, Douglas,; "Voice Encoding Methods for Digital Wireless Communications Systems", EE6302 Section 324, Fall, Southern Methodist University, (1997)

[9] Blelloch, Guy E.; "Introduction to Data Compression", Computer Science Department Carnegie Mellon University, (2001)

[10] Florian Metze, Hao Zhang, Yajie Miao, Member, IEEE, “Speaker Adaptive Training of Deep Neural Network Acoustic Models Using I-Vectors,” IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 23, NO. 11, NOVEMBER 2015