ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

FPGA IMPLEMENTATION FOR THE ROBUST FEATURE EXTRACTION FOR SPEECH RECOGNITION

Mohammad Javeed,
Assistant Professor, Dept. of ECE,
Sree Dattha Institute of Engineering and Science, India

Dr. M.Senthil Kumar,
Professor, Dept. of ECE,
Sree Dattha Institute of Engineering and Science, India

Abstract: Speaker recognition in noisy environments is a challenge when there is a mismatch in the data used for enrollment and verification. In this project, we propose a robust functional extraction system based on modulation filtration using two-dimensional (2D) autoregressive (AR) models. The first step is AR modeling of sub-band temporal envelopes using the linear prediction on sub-band discrete cosine transforming (DCT) components. These sub-band envelopes are stacked together and used for another AR modeling step. The spectral envelope across the subbandles is approximated in this AR model, and cepstral functions are derived from which are used for speaker recognition. The use of AR models underlines the focus on the high-energy regions that are relatively well preserved near noise. The gradient filtering rate is controlled by the AR Model Order parameter. Experiments are conducted using noisy versions of NIST 2010 speaker recognition evaluation (SRE) data with a state of art speaker recognition system. The proposed system implemented in MATLAB and then implemented in XILINX and Modelsim.

Key words: Robust feature extraction, speech recognition, AR model, FPGA.

I. INTRODUCTION

Speech technology works reasonably in matched conditions but rapidly degrades when there is acoustic mismatch between the training and test conditions. Although multi-condition training can improve the performance [1] , realistic scenarios can benefit from more robustness without requiring training data from the target acoustic environment. In this paper, we develop a feature extraction scheme which attempts to address robustness in noisy and reverberant environments.

Automatic speech recognition technology has a high potential for improving the learning experience of students in an educational setting. Some of the key theoretical areas involved in developing automatic speech recognition systems for educational use; namely the applications of the technology in education, prominent feature extraction and noise cancellation techniques used with audio speech data as well as some of the recent neural network based machine learning models capable of keyword spotting or continuous speech recognition shown in [2]. Authors have discussed some of the traditional feature extraction techniques that are commonly used in the areas of language identification speech recognition, and speaker verification, and their pros and cons in [3]. Due to the nonlinear nature of speech, LPC are not generally used for speech estimation. It was discussed that the most frequently used feature extraction techniques are MFCC, LPC and PLP in the areas of speaker verification and speech recognition applications but recently hybrid features are overcoming the traditional features [3].

In the past, various feature processing techniques like spectral subtraction, Wiener filtering and missing data reconstruction have been developed for noisy speech recognition applications. Feature compensation techniques have also been used in the past for speaker verification systems (feature warping , RASTA processing and cepstral mean subtraction (CMS))[4].With noise or reverberation, the

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

low energy valleys of speech signal have the worst signal to noise ratio (SNR), while the high energy regions are robust and could be well modeled[5][6]. In general, an autoregressive (AR) modeling approach represents high energy regions with good modeling accuracy. The AR modeling approach of signal spectra is widely used for feature extraction of speech . The AR modeling of Hilbert envelopes have been used with similar goals of preserving peaks in sub-band temporal envelopes and has been successfully applied for speaker verification. 2- D AR modeling was originally proposed for speech recognition by alternating the AR models between spectral and temporal domains[7].

In this paper, we extend the previous approach on two dimensional AR modeling with a modulation filtering framework. Long segments of the input speech signal are decomposed into sub-bands and linear prediction is applied on the sub-band discrete cosine transform (DCT) components to derive Hilbert envelopes [8]. The sub-band envelopes are stacked together to form a time-frequency description and a second AR model is applied across the sub-bands for each short-term frame (25 ms with a shift of 10ms)[9][10]. The output of the second AR model is converted to cepstral coefficients and used for speaker recognition. Modifying either of the AR models, time domain one or the frequency domain one, represents in effect a ratescale (time-frequency) modulation filtering. The time domain AR model does the rate filtering and the frequency domain AR model does the scale filtering, similar to the approaches discussed[11]12]. Experiments are performed on core conditions of NIST 2010 SRE data with various artificially added noise and reverberation[13].

II. METHODOLOGY

The block schematic for the proposed feature extraction is shown in Fig. 1. Long segments of the input speech signal (10s of non-overlapping windows) are transformed using a discrete cosine transform . The full-band DCT signal is windowed into a set of 96 over-lapping linear sub-bands in the frequency range of 125-3700 Hz. In each sub-band, linear prediction is applied on the sub-band DCT components to estimate an all-pole representation of Hilbert envelope. This constitutes the temporal AR modeling stage. The FDLP envelopes from various sub-bands are stacked together to obtain a two-dimensional representation as shown in Fig. 1.

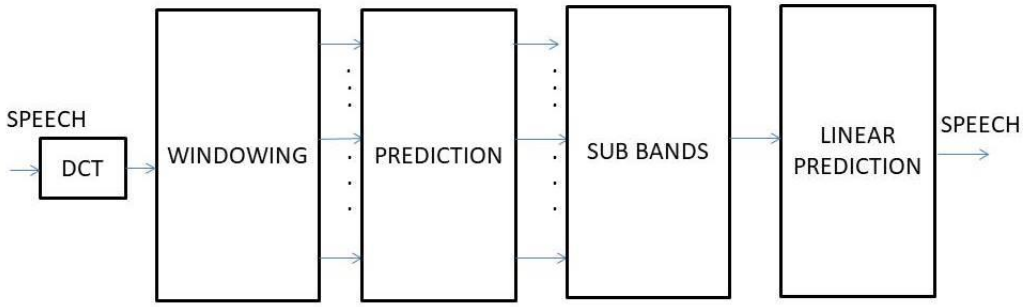


Figure 1: Block schematic of the proposed feature extraction using AR Models.

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

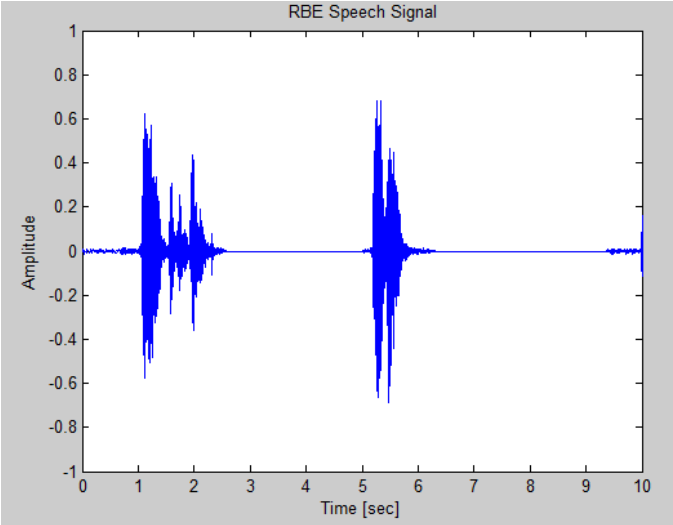


Figure 2. Robust feture extraction in RBE signal.

(25ms with a shift of 10ms). The output of the integration process provides an estimate of the power spectrum of signal in the short-term frame level. The frequency resolution of this power spectrum is equal to the initial sub-band decomposition of 96 bands[14]. These power spectral estimates are transformed to temporal autocorrelation estimates using inverse Fourier transform and the resulting autocorrelation sequence is used for time domain linear prediction (TDLP). We derive 13 cepstral coefficients from the all-pole approximation of the 96 point shortterm power spectrum. The delta and acceleration coefficients are extracted to obtain 39 dimensional features.

A. Rate-Scale filtering using AR Models

A temporal modulation filter is referred to as a rate filter and a spectral modulation filter is referred to as a scale filter. In the proposed feature extraction framework, the AR modeling process represents a filter impulse response, whose frequency response ("time response" in the case of the temporal AR filter) can be controlled by the model order. A lower model order represents more smoothing in a given domain , while the higher model captures finer details . Thus, various streams of spectrographic representations can be generated from the proposed framework using different choices of model order for temporal and spectral AR models as shown in Fig. 2. The low-rate lowscale representations represent broad energy variations in the signal as seen in Fig. 2 . The other configuration using higher order for the AR models is shown in Fig. 2 where more details about the various events in the spectrogram are evident. A higher order could also mean that such AR models may carry information about noise or reverberation artifacts that is present in the finer details of the spectrogram in its spectral or temporal directions. In addition to the configurations shown in Fig. 2, other possibilities include a lower model order for temporal AR model with a higher order for the spectral AR model and vice-versa. Thus, various feature streams which differ in the extent of modulations can be derived from the spectro-temporal AR model framework. In Sec. 3, we provide some experiments showing the effect of model order on the speaker recognition performance [15].

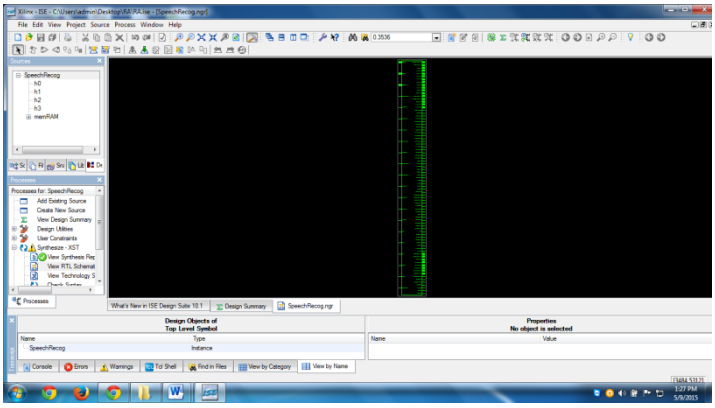
B. Robustness to Noise

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

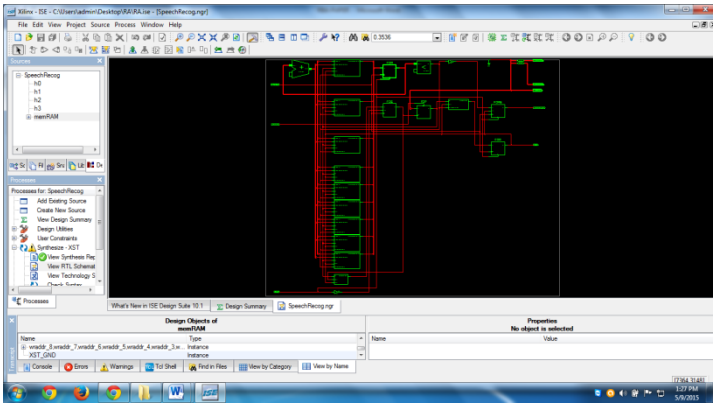
When a speech signal is corrupted with noise or reverberation, the valleys in the sub-band envelopes are dominated by noise. Even with moderate amounts of distortion, the low-energy regions are substantially modified and cause acoustic mismatch with the clean training data. Since the AR modeling tends to fit the high energy regions with good accuracy. This is illustrated in Fig. 2 where we plot a portion of clean speech signal, speech with additive noise (babble noise at 10 dB SNR) and speech with artificial reverberation (reverberation time of 300 ms). The spectrographic representation obtained from mel frequency representation is shown in the second panel and the corresponding representation obtained from spectro-temporal AR models is shown in the bottom panel. In comparison with the mel spectrogram, the representation obtained from AR modeling emphasizes the high energy regions[16]. Thus, such a representations can be more similar for the clean and the noisy versions of the same signal. This is desirable and contributes to improved robustness when these features are used for speaker recognition in noisy environments.

III. SIMULATION RESULTS

The figures 3 (a), (b), (c) shows the RTL diagrams for the proposed design. Executed in Xilinx and the code has been written in Verilog. Figure 4 (a) and (b) shows the waveforms generated in the modelsim.

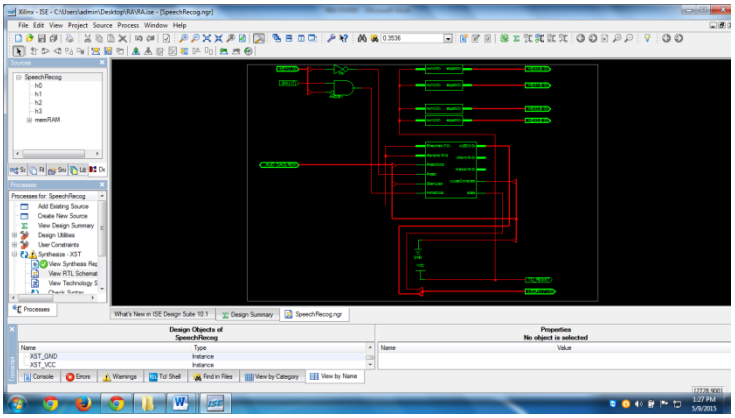


(a)



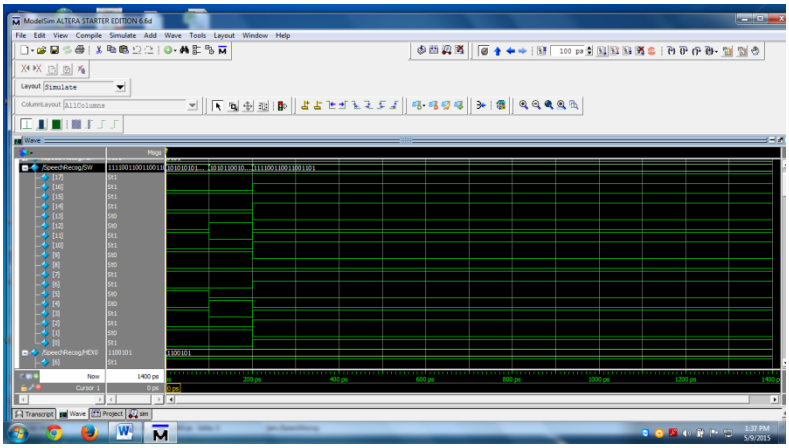
(b)

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

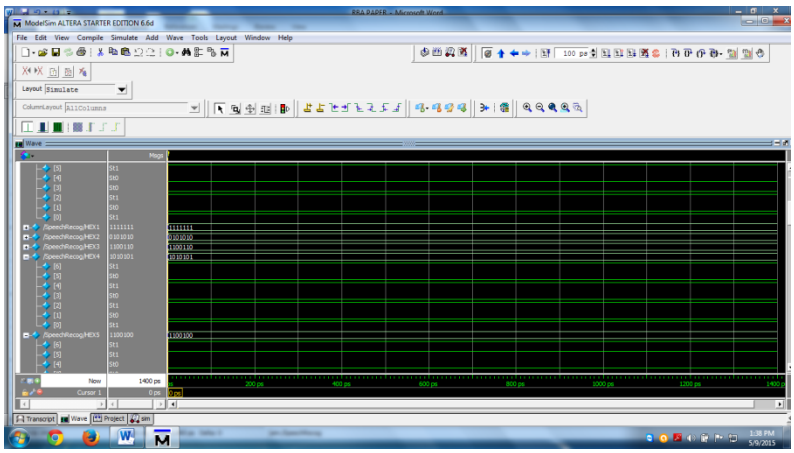


(c)

Figure 3. RTL diagrams



(a)



(b)

Figure 4. Simulation wave forms generated in Modelsim.

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

IV. CONCLUSION

In this paper, we have proposed a two-dimensional autore regressive model for robust speaker recognition. An initial temporal AR model is derived from long segments of the speech signal. This model gives Hilbert envelopes of sub-band speech, which are integrated into short-term frameworks to achieve power spectral estimation. The estimates are used for a spectral AR modeling process, and the output prediction coefficients are used for speaker recognition. Various experiments are conducted with noisy test data on the NIST 2010 SRE, where the proposed features provide significant improvements. These results are also validated using a large speaker recognition data set from BEST. The results are promising and encourage us to pursue the problem of joint 2-D AR modeling instead of a separable time and frequency linear prediction schemes adopted in this project. Implemented the AR based RBE in FPGA using Xilinx.

REFERENCES

[1]. Ming, J., Hazen, T.J., Glass, J.R. and Reynolds, D.A., "Robust Speaker Recognition in Noisy Conditions", *IEEE Tran. on Audio Speech Lang. Proc.*, Vol 15 (5), 2007, pp. 1711 - 1723.

[2]. Phillip Blunt ; Bertram Haskins," A Model for Incorporating an Automatic Speech Recognition System in a Noisy Educational Environment", 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC).

[3]. Usha Sharma, Sushila Maheshkar, A.N.Mishra, "Study of Robust Feature Extraction Techniques for Speech Recognition System", 2015 1st International conference on futuristic trend in computational analysis and knowledge management (ABLAZE 2015)

[4]. Boll, S.F.,"Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 27 (2), Apr. 1979, pp. 113-120.

[5]. "ETSI ES 202 050 v1.1.1 STQ; Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms", 2002.

[6]. Cooke, M., Morris, A., Green, P., "Missing data techniques for robust speech recognition", *Proc. ICASSP*, 1997, pp. 863-866.

[7]. Pelecanos, J. and Sridharan, S., "Feature warping for robust speaker verification", *Proc. Speaker Odyssey 2001 Speaker Recognition Workshop*, Greece, pp. 213-218, 2001.

[8]. Hermansky, H. and Morgan, N., "RASTA processing of speech," *IEEE Trans. on Speech and Audio Process.*, Vol. 2, pp. 578-589, 1994.

[9]. Rosenberg, A.E., Lee, C. and Soong, F.K., "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification," in *Proc. ICSLP*, pp. 1835-1838, 1994.

[10]. Davis, S. and Mermelstein, R., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 28 (4), Aug. 1980, pp. 357-366.

[11]. Guruprasad, S., "Significance of processing regions of high signal-to-noise ratio in speech signals", PhD Thesis, 2011.

[12]. Atal, B.S., Hanauer, L.S, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoust. America*, Vol 50 (28), 1971, pp. 637-655.

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

[13]. Makhoul, J., "Linear Prediction: A Tutorial Review",in *Proc. Of the IEEE*, Vol 63(4), pp. 561-580, 1975.

[14]. Hermansky, H., "Perceptual Linear Predictive (PLP) Analysis of Speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738-1752, 1990.

[15]. Ganapathy, S., Pelecanos, J. and Omar, M.K., "Feature Normalization for Speaker Verification in Room Reverberation", *Proc. ICASSP*, 2011, pp. 4836-4839.

[16]. Sriram Ganapathy, *Member, IEEE*, Sri Harish Mallidi, *Student Member, IEEE*, and Hynek Hermansky, *Fellow, IEEE,* "Robust Feature Extraction Using Modulation Filtering of Autoregressive Models", IEEE/acm transactions on audio, speech, and language processing, vol. 22, no. 8, august 2014.