INTEGRATION, the VLSI journal xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

INTEGRATION, the VLSI journal

journal homepage: www.elsevier.com/locate/vlsi

An efficient hardware architecture for detection of vowel-like regions in speech signal

Nagapuri Srinivas, Gayadhar Pradhan*, Puli Kishore Kumar

Department of Electronics and Communication Engineering, National Institute of Technology Patna, India

ARTICLE INFO

Keywords:
Non-local slope difference
Non-linear mapping
Vowel-like regions
Zero frequency filtering
Hardware architecture FPGA

ABSTRACT

Vowel-like regions (VLRs) in a speech signal include vowel, semivowel and diphthong sound units. In the existing VLRs detection methods, front-end speech parameterization have been done by complex algorithms. Those approaches require more hardware and hence delay the process. To address this issue, a simple and robust signal processing approach and it's hardware architecture is proposed for discriminating VLRs in the speech signal. In the proposed approach, non-local slope difference (*NSD*) at each time instant is computed by processing the speech signal through a single pole filter. The *NSD* is then averaged over an analysis frame and non-linearly mapped using negative exponential to reduce the fluctuations present in the input speech signal. The non-linearly mapped averaged *NSD* (*NL-ANSD*) is used as the front-end feature for discriminating VLRs. The *NL-ANSD* exhibits significantly sharp transition at the starting and ending points of the VLRs. The regions wherein the proposed feature exhibits significant transition and attains lower magnitude for a considerable duration of time are hypothesized as the VLRs. The proposed approach is very simple and requires significantly less hardware when compared with the existing zero-frequency filtering (ZFF) based methods. On the other hand, the proposed approach outperforms the existing ZFF based approaches for the task of detecting VLRs in clean as well as noisy speech signals. The hardware architecture of the proposed approach is verified by implementing it on the Nexys video *Artix* − 7(*XC7A*200T − 1*SBG*484C) field-programmable gate array (FPGA) trainer board for multimedia applications using Xilinx system generator-2016.2.

1. Introduction

In a given speech signal semivowels and diphthongs are more similar to the vowels when compared to the other sound units [1,2]. In this paper, vowel, semivowel, diphthong sound units are collectively termed as vowel-like regions (VLRs) [1,2]. During production of the speech signal, most significant excitation of the vocal-tract system takes place within these regions [1,3]. The VLRs are also produced by the vibration of vocal folds. Consequently, VLRs are periodic, high energy and long duration sound units [3]. Due to high energy, these regions are less affected in the noisy speech signals when compared to the other sound units. Analysis of the VLRs provides a better estimation of the excitation source information as well as the frequency response of the vocal-tract system [1,2]. In the earlier reported works, knowledge of the vowels and VLRs have been used for building effective speaker recognition systems [1,2,4–6]. The knowledge of vowel is also explored for the development of speech based applications, like keyword spotting [7], speech segmentation [8,9], dialect classification [10], emotion classification [11,12] and prosody modification [13–15]. Vowels have also been employed in the analysis and detection of different disorders like dysphasia [16], physiological and neurological disorders [17], Parkinson's disease [18], etc., using speech data.

1.1. Existing methods for discriminating VLRs

Extracting relevant front-end acoustic features which are discriminative for VLRs and less affected under noisy test conditions is very challenging. The energy and periodicity of the speech signal can be used as features for detecting VLRs. Short-term energy and periodicity information directly computed from the speech signal varies significantly depending on the signal to noise ratio (SNR) [19–21]. In the last few years, only for the detection of vowel onset point (VOPs)/vowels, several front-end speech parameterization features and statistical modeling methods have been proposed [1,2,19,22–27]. In most of the cases, the existing methods fail to discriminate the semivowels and vowels due to their similarities in temporal and spectral domains [1,26,27]. If the semivowels are grouped with the vowels, the existing vowel detection methods can also be used for the detection of VLRs.

* Corresponding author.
E-mail addresses: ns@nitp.ac.in (N. Srinivas), gdp@nitp.ac.in (G. Pradhan), pulikishorekumar@gmail.com (P.K. Kumar).

https://doi.org/10.1016/j.vlsi.2018.07.005
Received 8 January 2018; Received in revised form 26 May 2018; Accepted 18 July 2018

The earlier reported works on detecting vowels/VLRs may be categorized into two group of approaches. In the first group, front-end speech parameterization is done in such a way so as to enhance the distinct nature of excitation source and vocal-tract system within the VLRs [19,22,23,28,29]. These front-end features include the difference in energy of each of the peaks and their corresponding valleys in the magnitude spectrum [28], zero-crossing rate, energy and pitch information of the speech signal [24], wavelet scaling coefficients of the speech signal [30], Hilbert envelope of the linear prediction (LP) residual [31], spectral peaks, modulation spectrum energies [22], rate of change of excitation strength extracted from the zero frequency filtered (ZFF) speech signal [1,2], spectral energy present in the glottal closure regions [23] and uniformity of the epoch intervals in vowels [19]. In Ref. [21], non-local estimation of the speech signal was done to separate the high energy voiced regions from the unvoiced regions. Then, the cumulative sum of the short-term magnitude spectrum is computed for discriminating vowels. Several vocal-tract and excitation source features have also been combined to represent the complementary information present in the vowels [1,2,19,22,27].

In the second group, different acoustic modeling techniques have been explored to develop classifiers that could differentiate between VLRs and non-VLRs [24–27]. For the development of effective classifiers, the statistical modeling methods like Hierarchical neural network, multilayer feed-forward neural networks and auto-associative neural networks have been used [24,25]. Recently, the hidden Markov model (HMM) was explored for the detection of vowels [26,27]. For modeling the observation densities of the HMM states, different techniques like Gaussian mixture modeling (GMM), subspace GMM (SGMM) [32] and deep neural network (DNN) [33] were explored. The approaches based on statistical modeling provide improved performance compared to those based on explicit signal processing. At the same time, their performance varies with the choice of statistical modeling techniques and kind of front-end speech parametrization method employed [26,27]. Further to this, performance also varies depending on the amount and nature of the training data available to learn the model parameters.

The approaches based on statistical modeling require voice activity detection, extraction of multidimensional features, feature normalization at the front-end and complex acoustic modeling during training [26,27]. At the time of testing, a given test utterance is decoded using the trained acoustic models. It is very similar to the development of a speech recognition system where the possible output classes are VLR, non-VLRs and silence. Developing such a system on integrated chip (IC) is very complex and requires a significant amount of hardware. On the other hand, explicit signal processing approaches do not require a training phase. Detection of the VLRs can be done for real-time applications using discriminative features only. For real-time applications, combination of multiple features and evidences make the system complex and delay the process. Therefore, the hardware realization demands a single robust feature for an effective detection of VLRs.

1.2. Contributions of the present work

Out of the different front-end speech parameterization methods proposed in the literature, as a single dimensional feature, the information extracted from the ZFF of speech signal provides a relatively better performance [1,2,29]. The ZFF based approaches are also robust to the environmental noises. However, the hardware design for the ZFF based VLRs detection methods is complex and also requires more hardware. Therefore, a simple method which effectively discriminate VLRs from non-VLRs is desirable for efficient hardware implementation.

Field-programmable gate arrays (FPGAs) are re-configurable hardware integrated chips (ICs) that can be programmed to implement different digital systems. The field programmable functionality implemented in FPGAs is defined by the user rather than being fixed at the time of manufacturing. Hence, these are more suitable for verifying digital system hardware architectures and prototyping digital signal processing (DSP) algorithms. FPGAs have been successfully applied in different domains like telecommunications, robotics, pattern recognition and infrastructure monitoring [34–38]. As far as our knowledge is concerned, VLRs detection algorithms are not yet been implemented on hardware.

In this paper, we have proposed a simple VLRs detection method and it's hardware architecture. The performance of the proposed approach is compared with two existing ZFF based approaches simulated using MATLAB [2,23]. The experimental results presented in this work shows that, for clean as well as noisy speech signals, the proposed approach outperforms the existing ones. On the other hand the hardware requirement for the proposed method is significantly less when compared to the ZFF based approaches.

To summarize, the novel contributions of the work presented in this paper are as follows:

i) A simple and highly effective signal processing approach for extracting the non-local slope difference (*NSD*) is proposed. The average of *NSD* (*ANSD*) over an analysis frame happens to be significantly higher for the VLRs when compared with the non-VLRs, silence and noise regions.
ii) The *ANSD* at each time instant is non-linearly mapped (*NL-ANSD*) to suppress the fluctuations present within the VLRs. The NL-ANSD is equally discriminative at the starting and ending points of the VLRs. The proposed approach address the issue of detecting starting and ending points of VLRs in a single algorithm which remained as a very challenging task [2,20].
iii) An efficient hardware architecture is designed for the proposed method and is implemented on FPGA for real-time detection of VLRs.

The rest of the paper is organized as follows: Section 2 discusses the development of two existing VLRs detection algorithms using features extracted from the ZFF speech signal. Section 3 presents the proposed method for VLRs detection. Proposed method is compared with the explored VLRs detection methods in Section 4. Hardware architecture design of the proposed method and it's implementation on FPGA is presented in Section 5. Finally, the paper is concluded in Section 6.

2. Development of ZFF-based VLRs detection methods

The discontinuities due to impulse like excitation is spread over all frequencies including the zero frequency. The ZFF output of the speech signal preserves the energy only around the zero frequency and filter out all other frequency components [39]. The vocal-tract response around the zero frequency is negligibly small. Therefore, the output of ZFF consists mainly of excitation source information. Initially, the ZFF was implemented using IIR realization [39]. In that realization, cascaded of two zero frequency resonators and two detrenders were used. The output of the ZFF filter is a third degree growing polynomial with respect to the time. Consequently, bit precision required for the hardware implementation is proportional to the duration of speech signal. To overcome the problem of marginal stability in IIR realization, a stable FIR realization of ZFF is proposed in Ref. [40] by employing the inherent pole-zero cancellation. For a given input speech signal ($x(n)$), in the FIR realization ZFF signal is obtained as follows [40]:

$$y(n) = \frac{-1}{2N+1}\{S_1, S_2, ...,S_{N-1}, S_N, S_{N-1}, ...,S_2, S_1\} * x(n+N) \quad (1)$$

where $y(n)$ and S_N represents the ZFF speech signal and the sum of N natural numbers ($S_N = N(N+1)/2$), respectively. The $*$ symbol represents the convolution operation.