CHAPTER

8

VERILOG IMPLEMENTATION OF VOICE ACTIVITY DETECTION

Ng Boon Khai, Muhammad Mun'im Ahmad Zabidi, and Shahidatul Sadiah Abdul Manan

8.1 INTRODUCTION

Voice Activity Detection (VAD) is a critical front-end processing stage for all speech-related applications. It detects the presence or absence of human speech in the presence of other sounds (Yoo et al., 2015). During the non-speech phase of an audio session, VAD deactivates some operations or circuits to save energy and prevent wasted coding or network transmission of packets that do not contain voice (Figure 8.1). Voice over Internet Protocol (VoIP) programs use VAD to reduce compute power and network bandwidth. To be effective, VAD circuits should have high accuracy while consuming very little power. A poorly performing VAD front-end hurts the performance of speech-processing systems (Ali & Talha, 2018).

Human voice contains important information such as energy, pitch, loudness, prosody, and voice quality, biological information, and paralinguistic information (e.g., social status, personality traits, and emotional state of the speaker) (Doehring & Bartholomeus, 1971; Papcun et al., 1989; Van Dommelen, 1990; Zhang, 2016). The larynx, often known as the voice box, produces the human voice by housing the vocal cords, which are used to modulate pitch and loudness and

are necessary for phonation (Zhang, 2016). The structure of the larynx, as illustrated in Figure 8.2, as well as its physical properties and modes of vibration, affect the kind and quality of sound generated.

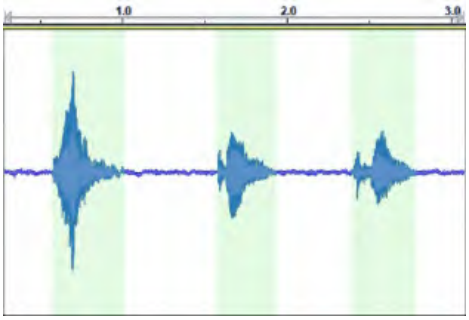

Figure 8.1 Voice activity detection outputs a true signal when human voice detected

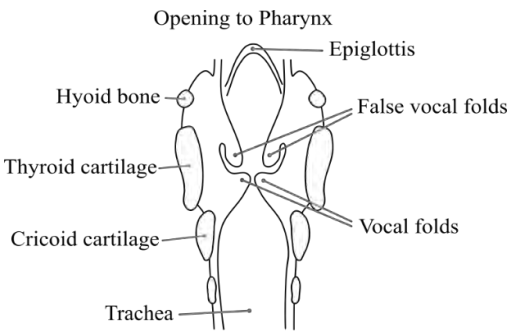

Figure 8.2 The larynx

8.2 VOICE ACTIVITY DETECTION ALGORITHMS

The objective of our work was to implement a VAD proof-of-concept rapidly. This can be done at the register transfer level using the Verilog hardware design language. We have access to Intellectual Property (IP) blocks which can be used to execute the algorithm by simulation. To achieve this target, we investigated several possible algorithms.

Simple VAD algorithms used short-term energy or zero-crossing rate to detect voice. The energy-based approach to VAD is shown in Algorithm 8.1:

Algorithm 8.1 Energy Based VAD Algorithm

Input: sample sound

Output: audio data labelled as voice

1: convert the sample sound to mono
2: **while** move a window of 20ms along the audio data
   (1) calculate the ratio between the energy of speech band (e.g., 300 to 3000 Hz) and total energy of the window
   (2) **if** the ratio is higher than a predefined threshold (e.g., 60%)
   (3) label it as voice
   (4) **else**
   (5) continue to next window
   (6) **end if**
   (7) **end while**
3: Return

As the noise level rises, the performance of simple algorithms frequently suffers (Graf et al., 2015). To address this issue, robust acoustic features were used including statistical features (Ramirez et al., 2005; Sohn et al., 1999; Xu et al., 2015), autocorrelation function- based features (Kristjansson et al., 2005), entropy (Xu et al., 2015), long-term signal variability (Ghosh et al., 2010) and cepstral features (Haigh & Mason, 1993). Integrated audio-visual signals have been shown to give good performance but the methods necessitate the use of extra sensors, which are costly and cumbersome (Rivet et al., 2006). Finally, machine learning-based approaches give good performance at the expense of complexity (Price et al., 2017; Shin et al., 2010).

Voice Activity Detection (VAD) hardware implementations range from Field Programmable Gate Array (FPGA) to Application Specific Integrated Circuit (ASIC) (Jung et al., 2010; Meoni et al., 2018; Oukherfellah & Bahoura, 2014; Price et al., 2017). For efficient voice discrimination in non-stationary noise, the spectrum subtraction approach and short-time energy have been applied on FPGA (Oukherfellah & Bahoura, 2014). Real-time performance can be ensured by using signal periodicity and energy- based parameters (Jung et al., 2010). Because they are computationally tractable for real-time applications, short term features like spectral fitness and short-term