Stable Implementation of Voice Activity Detector Using Zero-Phase Zero Frequency Resonator on FPGA

Syed Abdul Jabbar, Purva Barche, Krishna Gurugubelli, **Syed Azeemuddin** and Anil Kumar Vuppala

Speech Processing Laboratory, LTRC, KCIS

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

International Institute of Information Technology, Hyderabad, India.

Real Time Communications Conference & Expo

2023 IEEE RTC

Outline

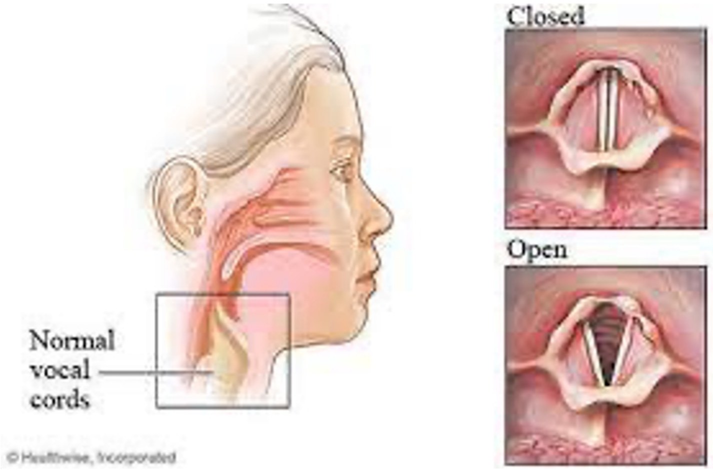
- Introduction
- Current State of Art
- Proposed Methodology
- Data-set
- Result
- Conclusion

Introduction

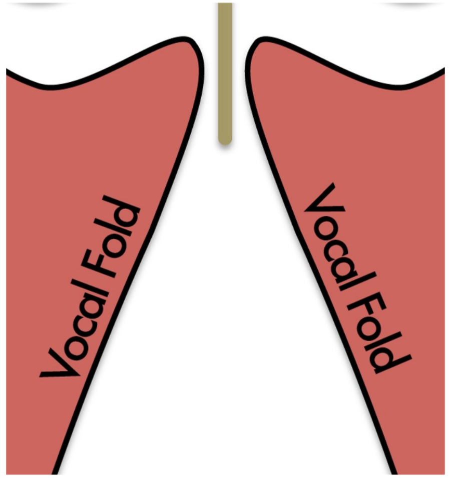
- Voice Activity Detection (VAD) – detect regions of significant vocal fold vibrations
- Such regions of speech are generally referred to as voiced speech
- In voiced speech the term voiced regions is used to refer those regions where the vibrations of the vocal folds are strong
- Voiced regions exhibit specific spectral characteristics, with energy concentrated at regular intervals corresponding to the harmonics

Voiced speech

- Vocal folds are two bands of muscle inside voice box (larynx) that allow to make sounds
- Vocal folds are closed when speaking, so the air from the lungs presses between them to cause vibrations
- The vibration creates the sound of your voice
- When you inhale or exhale, your vocal cords open so air can flow freely



© Healthwise, Incorporated



Unvoiced speech

- The nonvoiced regions of speech include both silence (or background noise) as well as unvoiced speech
- Unvoiced regions of speech include segments that do not involve the vibration of the vocal cords, such as sounds produced by voiceless fricatives (like “f” and “s”) and stops (like “p” and “t”)
- In unvoiced regions of speech, the vocal cords remain apart, allowing air to pass through freely without causing vibration
- In unvoiced speech, the vocal cords remain separated, allowing for the production of sounds like whispers or hisses, which do not involve vocal cord vibration

Current State of The Art

- VAD is useful in various applications related to speech processing and communication systems
  - Speech Coding, Voice Controlled Systems, speech feature extraction
- Over the years several algorithms have been proposed for Voice Detection
- VAD are mostly based on the zero-crossing rate (ZCR), energy levels, formant shape, linear prediction coding (LPC) parameters
- But most of these measures for determining voicing are sensitive to noise

Current State of Art

- Statistical models used - Neural Network Models, Gaussian Mixture Models (GMM), or Hidden Markov Models (HMM)
- Statistical approaches are more popular in voice activity detection (VAD) algorithms used in speech coding applications
- But not evaluate the performance of detecting voiced and unvoiced regions
- However, recently, Zero Frequency Filter (ZFF) algorithm for the VAD has been proposed
- One of the key advantages of the ZFF is its robustness to noise

Zero Frequency Filter (ZFF)

- The ZFF algorithm depends on excitation source information instead of speech signal for the detection of voice activity
- ZFF method is effective, but the response of the system may grow or decay rapidly which leads to the stability issue
- As ZFF has four poles on unit circle, these repeated poles on the unit circle make the ZFF unstable
- To overcome the stability problem, Zero-Phase Zero Frequency Resonator (ZP-ZFR) algorithm is proposed

Zero-Phase Zero Frequency Resonator (ZP-ZFR)

- The ZP-ZFR algorithm is a stable version of ZFF
- Technique used to extract significant instants in speech signals
- Has infinite impulse response (IIR) filter which requires lower filter order
- Stability of ZP-ZFR is due to the poles placed inside the unit circle
- Ensures consistent and reliable performance of the VAD system

Hardware Implementation

- The speed, reliability, and real-time capabilities of hardware implementations make VAD valuable in industries such as telecommunications, automotive, and consumer electronics.
- For hardware implementation, unstable algorithms may yield unpredictable results
- A stable implementation guarantees consistent and accurate identification of speech segments

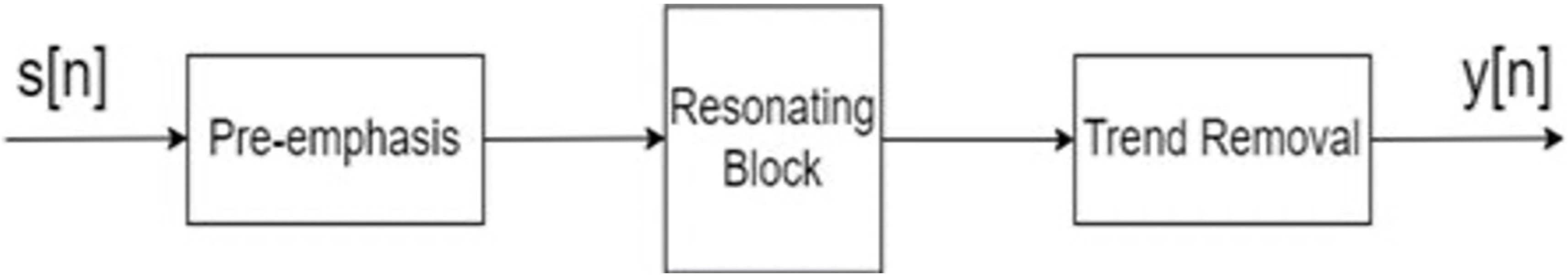
Verilog Implementation

- In this paper we implemented VAD on FPGA using ZP-ZFR algorithm, because of its stability
- A stability guarantees consistent and accurate identification of speech segments
- Up to our knowledge this approach is not yet implemented
- Hence, this paper proposes VAD based on ZP-ZFR algorithm using HDL (Verilog)

Proposed Methodology

- The ZP-ZFR algorithm plays a crucial role in accurately identifying voiced regions in a speech signal
- It can enhance the performance of VAD algorithm by providing clearer cues for distinguishing between voiced and unvoiced sounds
- The ZP-ZFR algorithm is designed to attenuate the vocal tract resonances while enhancing the excitation source of the speech signal
- It emphasizes the glottal activity, which is the vibration of the vocal folds responsible for generating the voiced portions of speech
- This resonator accentuates the low-frequency components of a signal

Block Diagram representation of ZP-ZFR



- The given speech signal passes through difference filter (pre-emphasis block) to remove low frequency fluctuations present in current signal
- This filtering stage helps to highlight the characteristics of the voiced segments in the signal

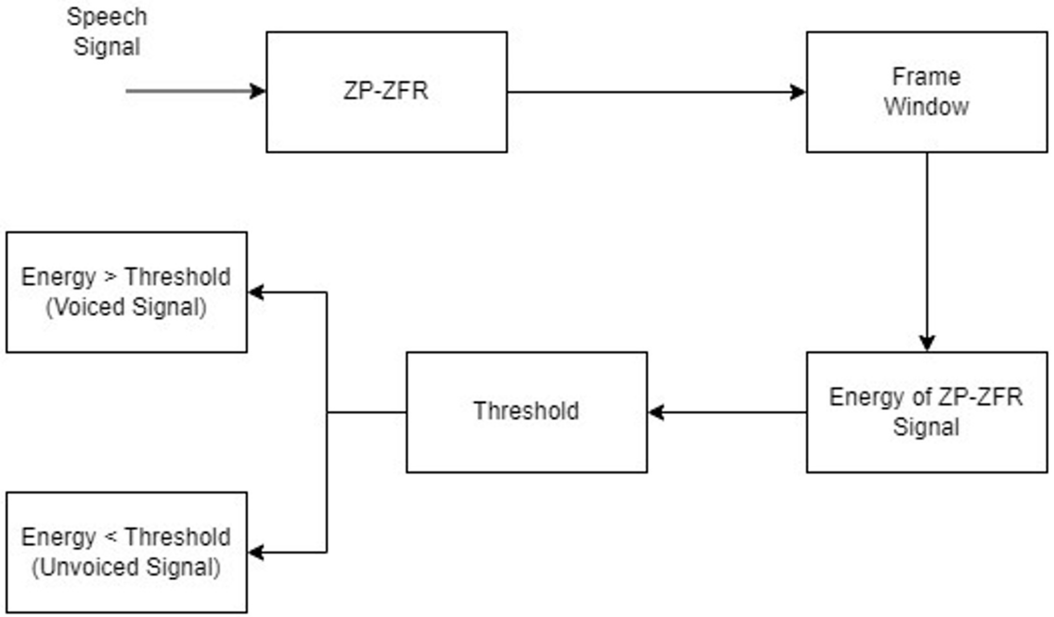
Proposed Methodology

- Then pre-emphasized speech signal passes through zero-phase zero frequency resonator block
- The purpose of passing the speech signal through ZP-ZFR is to highlight the impulse like excitation and to reduce the effects of all high frequency resonances present around the signal due to vocal tract resonance
- Then the resonating signal goes through trend removal block, which has high gain around the zero frequency which smears harmonics of speech signal
- This process helps in highlighting the discontinuities in the filtered signal due to impulse type of excitation

Proposed Methodology

- Trends in speech can arise due to factors such as variations in vocal tract shape, speaker-dependent characteristics, or environmental factors
- By removing these trends, these algorithms can enhance the accuracy of subsequent processing steps and improve the detection and analysis of specific speech features
- Generally, the fundamental period is in the range of 2.5 ms to 10 ms
- By considering the highest fundamental period, the analysis window for trend removal should be greater than 10 ms, to avoid the false alarms

Block Diagram representation of Voice Activity Detection


Speech Signal
ZP-ZFR
Frame Window
Energy of ZP-ZFR Signal
Threshold
Energy > Threshold (Voiced Signal)
Energy < Threshold (Unvoiced Signal)

- The figure depicts speech signal passes through ZP-ZFR algorithm
- Then this ZP-ZFR output signal is split into frames to be processed which has overlapping window

Proposed Methodology

- In this work, the frame size we used is 30ms and overlapping window of 20ms
- The energy calculation is applied to the short segments of the filtered output signal, which quantifies the strength or intensity of the signal
- To determine the presence of human speech activity, the calculated energy of the output signal is compared with a predefined threshold

Proposed Methodology

- The ZP-ZFR filtered signal exhibits high energy in the voiced regions due to significant contribution from the impulse-like excitation as compared to the unvoiced regions of speech
- If the energy exceeds the threshold value, it indicates the presence of voiced segments in the signal
- If the threshold value is greater than the energy, it indicates the presence of unvoiced segments
- Thus, the stable ZFF is implemented using Verilog on FPGA

Dataset

- The detection of voiced and unvoiced speech is evaluated on a subset of the TIMIT database
- The subset consists of 38 speakers
- The 24 male and 14 female, uttering 10 short sentences each
- All speech signals are recorded at a sampling rate of 16kHz

Results & Discussion

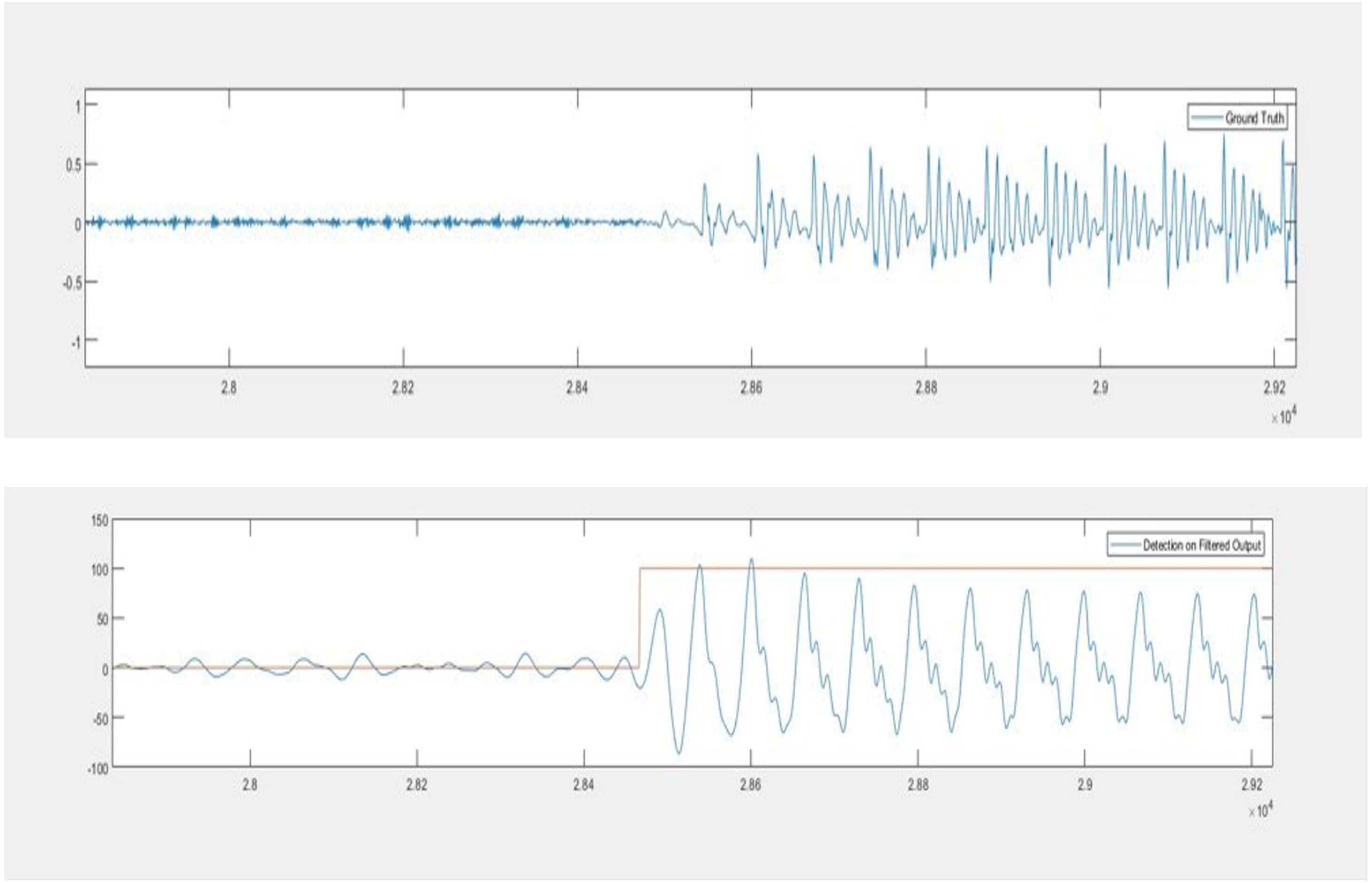
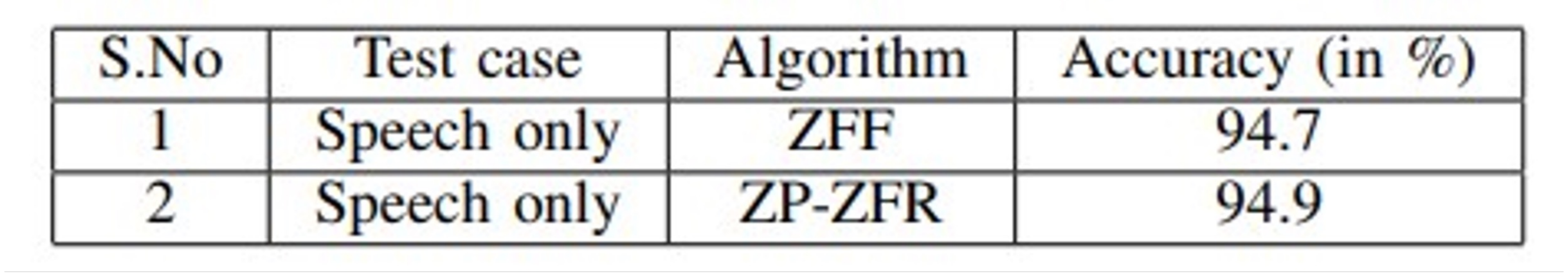


Figure: Voice Activity Detection using ZP-ZFR (a) A segment of speech signal (b) Voice Activity Detection using ZP-ZFR algorithm on segment of speech signal.

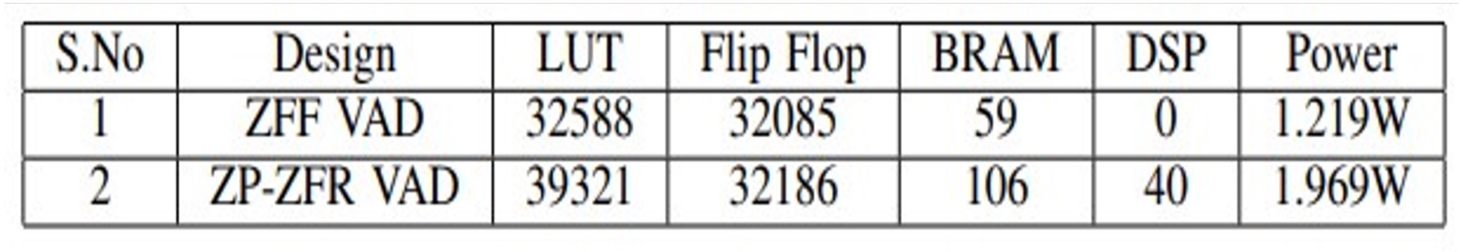
SOFTWARE PERFORMANCE OF VOICE ACTIVITY DETECTOR USING MATLAB



S.No	Test case	Algorithm	Accuracy (in %)
1	Speech only	ZFF	94.7
2	Speech only	ZP-ZFR	94.9

- The accuracy performance of ZP-ZFR based VAD is compared with the baseline ZFF based VAD method
- The result shows a slight difference in accuracy between the two approaches

HARDWARE UTILIZATION OF VOICE ACTIVITY DETECTOR USING VIVADO



S.No	Design	LUT	Flip Flop	BRAM	DSP	Power
1	ZFF VAD	32588	32085	59	0	1.219W
2	ZP-ZFR VAD	39321	32186	106	40	1.969W

- For the ZFF based VAD implementation, we can notice that the hardware utilization is less
- Upon comparing the two algorithms on hardware, it is evident that the ZPZFR based VAD exhibits more hardware utilization
- Despite the higher hardware and power utilization, we chose to implement the ZP-ZFR based VAD algorithm on hardware due to its stability and performance
- In future work, the power and hardware utilization of ZP-ZFR based VAD can be optimized

Conclusion

- ZFF and ZP-ZFR based VAD are implemented for identification of voiced locations
- The ZFF is a simple and most accurate technique among most of the other algorithms, but it is marginally stable
- The ZP-ZFR is stable implementation of ZFF with simple design and has better results than most of the ZFF parameters
- But the power and hardware utilization of ZP-ZFR based VAD costs more than the ZFF method
- ZP-ZFR's power and hardware consumption can be improved in future

References

- N. Dhananjaya and B. Yegnanarayana, “Voiced/nonvoiced detection based on robustness of voiced epochs,” IEEE Signal Processing Letters, vol. 17, no. 3, pp. 273–276, 2009.
- J. Jean-claude, “A study of endpoint detection algorithms in adverse conditions: Incidence on a dtw and hmm recognizer,” Eurospeach 1991, 1991.
- J. D. Hoyt and H. Wechsler, “Detection of human speech in structured noise,” in Proceedings of ICASSP’94. IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 2. IEEE, 1994, pp. II–237.
- L. Rabiner and M. Sambur, “Voiced-unvoiced-silence detection using the itakura lpc distance measure,” in ICASSP’77. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2. IEEE, 1977, pp. 323–326.
- A. P. Lobo and P. C. Loizou, “Voiced/unvoiced speech discrimination in noise using gabor atomic decomposition,” in 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03)., vol. 1. IEEE, 2003, pp. I–I.
- B. Atal and L. Rabiner, “A pattern recognition approach to voicedunvoiced-silence classification with applications to speech recognition,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 24, no. 3, pp. 201–212, 1976
- K. Gurugubelli and A. K. Vuppala, “Stable implementation of zero frequency filtering of speech signals for efficient epoch extraction,” IEEE Signal Processing Letters, vol. 26, no. 9, pp. 1310–1314, 2019.

THANK YOU