Paper ID: 10535

Unsupervised Voice Activity Detection by Modeling Source and System Information using Zero Frequency Filtering

Eklavya Sarkar[1,2], RaviShankar Prasad[1], Mathew Magimai Doss[1,2]

[1]Idiap Research Institute, Martigny, Switzerland [2]École polytechnique fédérale de Lausanne, Switzerland

Aims

- This paper investigates the potential of zero-frequency filtering for jointly modeling voice source and vocal tract system information, and proposes two approaches for Voice Activity Detection (VAD):

1. Demarcating voiced regions using a composite signal composed of different zero-frequency filtered signals.
2. Feeding the composite signal as input to the rVAD algorithm.

- These are compared with other supervised and unsupervised VAD methods in the literature, and evaluated on the Aurora-2 database across SNRs 20 to -5 dB.

Zero Frequency Filtering

- ZFF transforms the signal into filtered ones which contain f_0, F_1, and F_2 evidences.

$$x[n] = s[n] - 2x[n-1] + x[n-2] \tag{1}$$

$$y[n] = x[n] - \frac{1}{2N+1} \sum_{k=n-N}^{n+N} x[k]; \qquad N+1 \leq n \leq L-N. \tag{2}$$

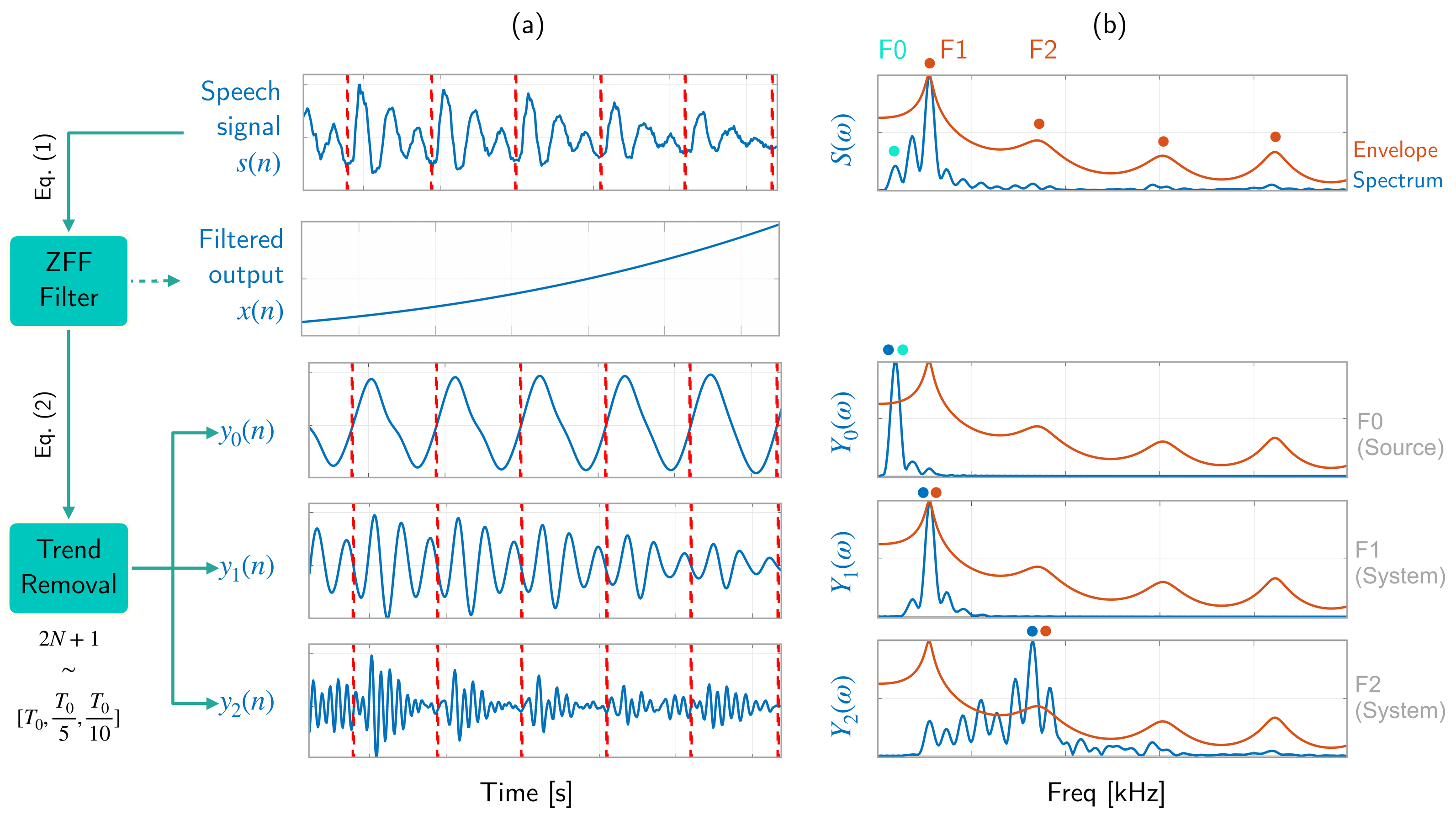
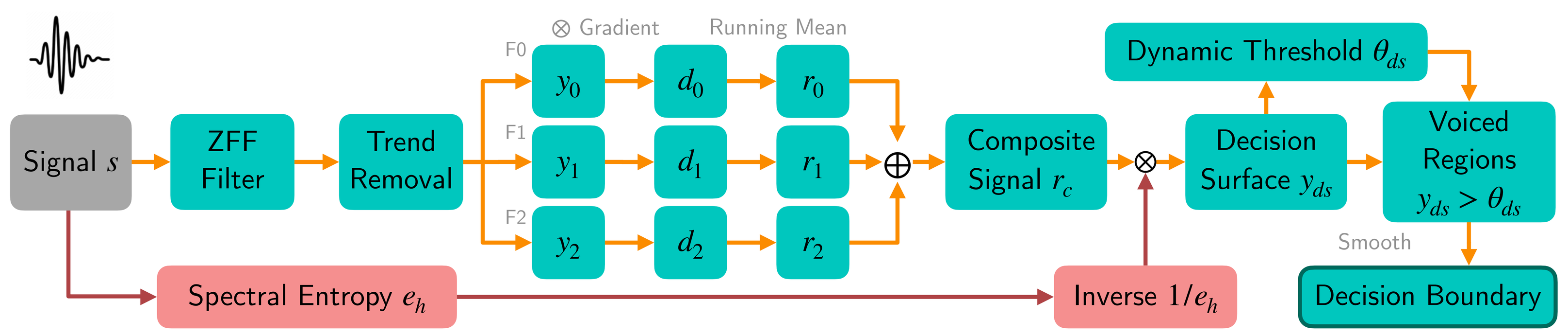


Figure 1. (a1) Speech signal. (a2) Filtered output. (a3–a4) ZFF signals $y_0(n)$, $y_1(n)$, $y_2(n)$. GCI locations (--). (b1) $S(\omega)$ (—) and its envelope (—). Formant peaks (•). Fundamental frequency peak (•). (b3–b4) $Y_0(\omega)$, $Y_1(\omega)$, $Y_2(\omega)$ (—), and respective peaks (•).

Proposed Method

- Pipeline of proposed method to derive a decision boundary for VAD:



Proposed Method

- Principal components of the ZFF-VAD technique:

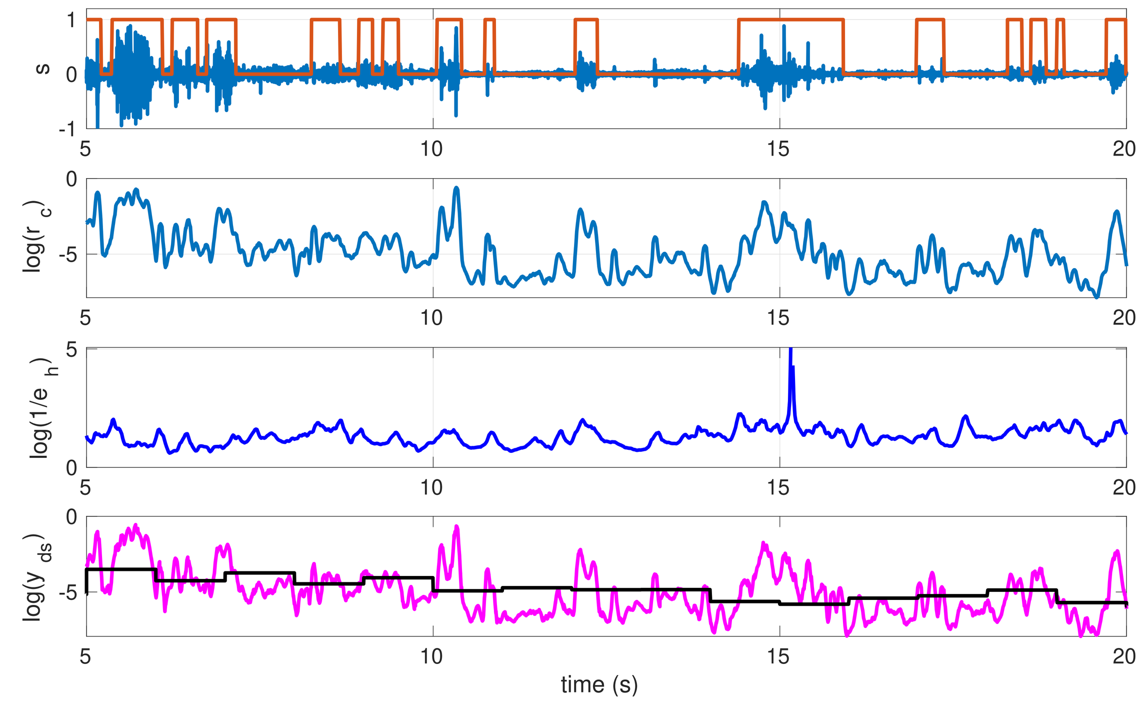


Figure 2. a) Naturally corrupted speech signal s and final decision boundary. b) Accumulated ZFF signals r_c c) Inverse spectral entropy $1/e_h$ d) Decision surface y_{ds} and dynamic threshold θ_{ds}.

Experimental Setup

Database, metrics, task:

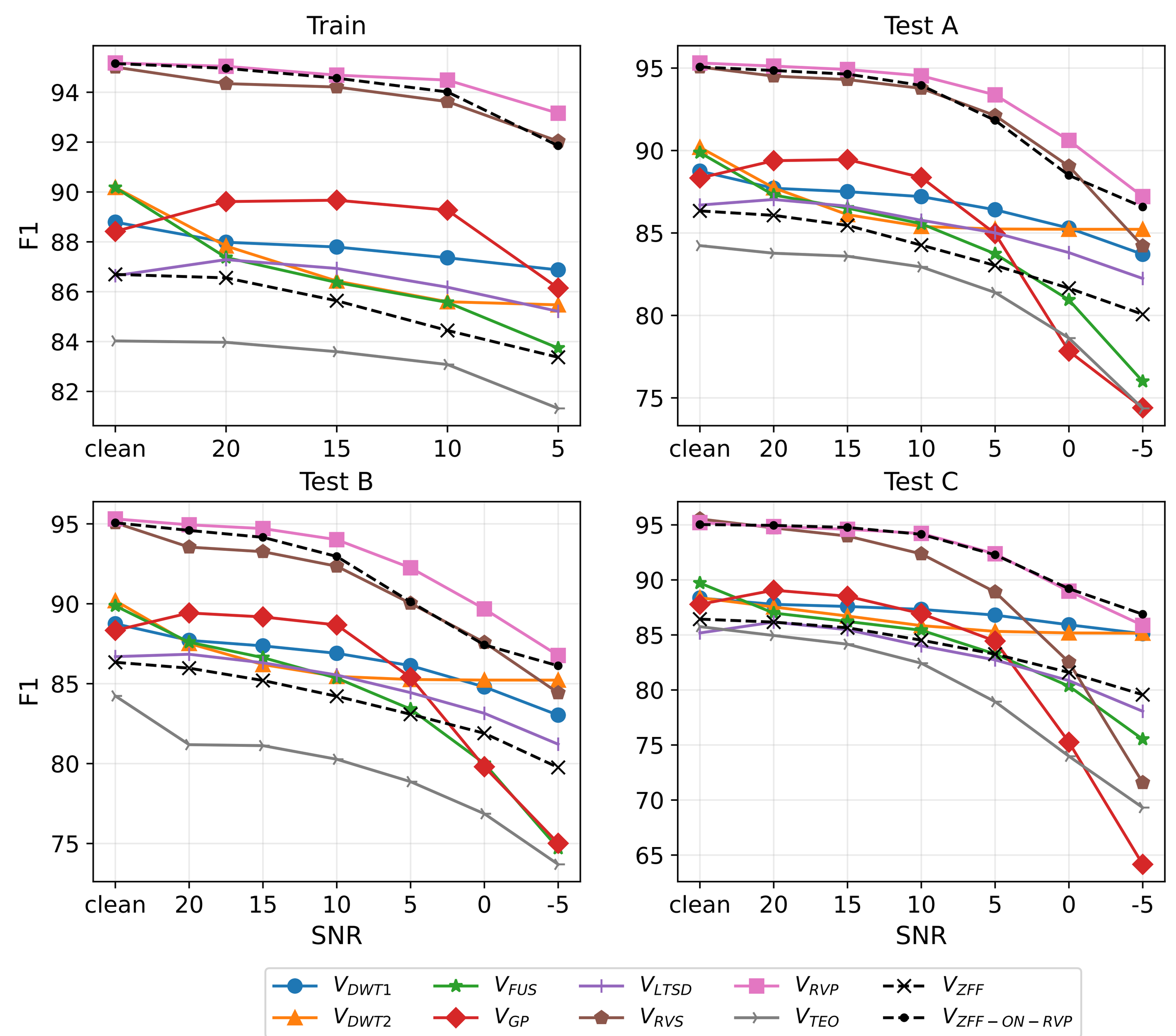
- Aurora-2
- F1-Score
- Binary classification

Baseline Methods:

- rVAD (V_{RVP})
- rVAD–Fast (V_{RVS})
- GPVAD (V_{GP})
- LTSD (V_{LTSD})
- Wavlet ($V_{\text{DWT1,2}}$)
- Fusion (V_{FUS})
- LSD (V_{LSD})
- TEO (V_{TEO})
- LSE (V_{LSE})

Results

- Classification performance of methods across all SNRs in different sets of Aurora-2:

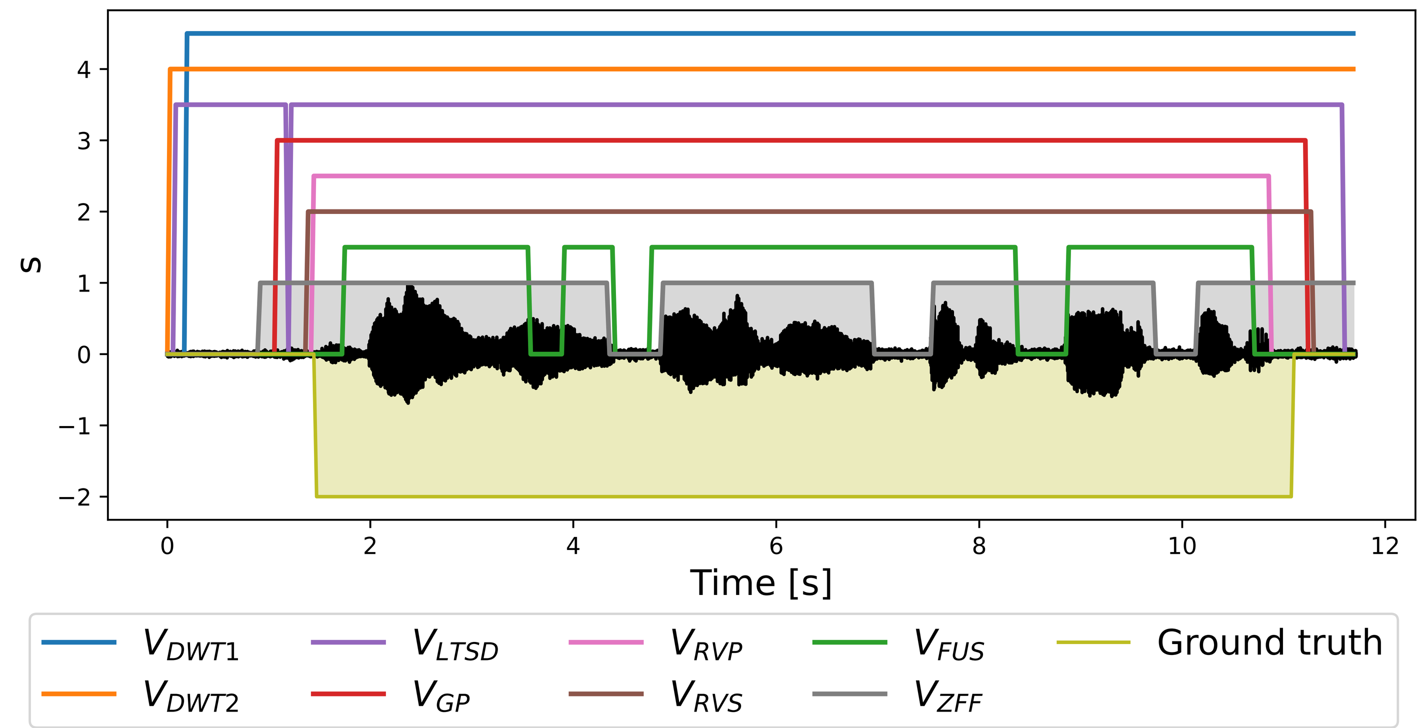


Analysis

- Standard deviation of the F1-scores of each method, across all SNRs of entire *Test* set.

V_{DWT}	V_{LSD}	V_{LTSD}	V_{ZFF}	V_{LSE}	V_{RVP}	$V_{\text{ZFF-ON-RVP}}$	V_{TEO}	V_{RVS}	V_{FUS}	V_{GP}
1.6	1.7	2.0	2.2	2.8	3.0	3.2	3.7	4.3	4.5	5.7

- Decision boundaries of all methods for a noisy speech sample (SNR = 10 dB):



➤ V_{ZFF} remains invariant to added interferences across a range of SNRs.
➤ V_{ZFF} segments the signal into significantly granular intervals than the other methods, as well as those given in the ground truth.

Conclusions

➤ VAD can be effectively performed with the proposed method i.e. by combining the ZFF filter outputs together to compose a composite signal carrying f_0, F_1, and F_2 related information, or else by passing the composite signal to another VAD.
➤ The composite signal, obtained by modulation of trend removal in the zero-frequency filtering, is an effective representation of speech characteristics, and can be used in conjunction with other VADs.

Future Work

- Model the composite signal using a raw waveform CNN and a self-supervised learning based modeling approach for robust supervised voice activity detection.
- The proposed framework could potentially be adapted to other types of audio signals, such as animal and birds vocalizations.

Acknowledgments

★ This work was funded by the Swiss National Science Foundation's NCCR Evolving language (grant agreement no. 51NF40_180888) and Towards Integrated processing of Physiological and Speech signals (TIPS) (grant agreement no. 200021_188754).