Hardware Implementations for Voice Activity Detection: Trends, Challenges and Outlook

S. Yadav, *Student Member, IEEE,* P.A.D. Legaspi, *Student Member, IEEE,*
M.S. Oude Alink, *Senior Member, IEEE,* A.B.J. Kokkeler, *Member, IEEE,* B. Nauta, *Fellow, IEEE*

***Abstract*—Voice Activity Detection (VAD) is a technique used to identify the presence of human voice in an audio signal. It is implemented as an always-on component in most speech processing applications. As speech is absent most of the time, this component typically dominates the overall average power consumption of the system (excluding microphone). The widespread usage in speech applications and the need for ultra low power VAD have led to a plethora of algorithms and implementations in the hardware domain, necessitating a comprehensive study and analysis to understand (real-time) requirements, different design parameters, testing strategies, but also to identify design trends, challenges and guidelines for future implementations and testing of VAD devices. A scoping review was conducted to identify the articles for hardware implementations of VAD from January 2010 - December 2021, the results of which are presented in this article. The results highlight a big design space being used for VAD along with a lack of standard testing methodology and usage of application-dependent performance metrics. An increased usage of filter-based feature extractors along with neural-network-based classifiers is observed. Due to lack of standardisation, no other trends can be established from the results. A set of rules and guidelines are therefore provided to facilitate the future development and benchmarking of VADs.**

***Index Terms*—Voice, Speech, Voice Activity Detection, Voice Detection, Review**

I. Introduction

ENERGY EFFICIENCY is an important requirement for portable edge devices like wearables and Internet-of-Things (IoT) nodes as they have small batteries or depend on energy harvesting circuits. An emerging application for portable devices is speech processing like Automatic Speech Recognition (ASR), Keyword Spotting (KWS), and Speaker Verification (SV). These applications are computationally intensive and power-hungry due to their 'always-on' nature. Therefore, they are typically preceded by a power gating component known as a Voice Activity Detector, interchangeably known as Voice Activity Detection (VAD), which is used to identify speech segments in a (noisy) audio signal, see Fig. 1, to reduce the overall energy consumption [1]–[3]. In the field of communication, e.g. mobile telephony, a VAD can avoid unnecessary processing and transmission of non-speech segments to save channel bandwidth as well [4].

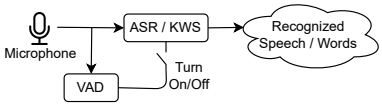

Fig. 1. Power gating functionality of VAD in speech processing applications like ASR and KWS (motivated from [5]–[8]).

Only when speech is detected, complex and power intensive components are turned on. The amount of time for which the components other than VAD stay off is highly application dependent. In case of a voice-activated switch (e.g. light control) which may be used only a few times a day, VAD keeps them off for $> 99.99\%$ of the time, whereas it can vary from 50% to 90% for ASR and SV systems [3]. This indicates that VAD, as an always-on block, can dominate the system's average power. In low power designs, the energy efficiency of the entire system significantly depends on how accurately VAD distinguishes between speech and non-speech segments.

VAD consists of multiple functional blocks working together to identify voice and noise segments, see Fig. 2. An active/passive electret condenser or MEMS microphone is placed at the input of the VAD to convert Sound Pressure Level (SPL) into equivalent electrical signals [9]. These electrical signals are then fed to an Analog Front End (AFE) which is typically made up of a Low-Noise Amplifier (LNA) and filters [6]. An Analog-to-Digital Converter (ADC) is placed after the AFE to provide input for the digital signal processing part to perform VAD. In case of analog VAD implementations, the ADC is placed either at the interface between the feature extractor (FEx) and the classifier or between VAD and other digital circuitry of the system [9] or it is absent [10]. In a digital VAD, the ADC is followed by the FEx and the classifier. The FEx extracts speech-related characteristics like short-time energy, short-time zero-crossing, Fast Fourier Transform (FFT) coefficients, etc. from the audio signal to compress it into a lower dimensional representation which are then fed into the classifier to differentiate between speech and non-speech segments. The output of the classifier, of which an example is given in Fig. 3, is then used to power-gate the complex components of the speech processing applications.

Covering the last 12 years (Jan 2010 - Dec 2021), we found several hardware implementations of VAD (either mea-

Manuscript received XX XX, 2022; revised XX XX, 2022; accepted XX XX, 2022. Date of publication XX XX, 2022; date of current version XX XX, 2022. This article was approved by (Associate) Editor ABCDEFG. This publication is part of the project Analog Approximate Accelerators (AAA) with project number 17703 of the research programme OTP which is (partly) financed by the Dutch Research Council (NWO).

S. Yadav and A.B.J. Kokkeler are with the Radio Systems (RS) group of University of Twente, The Netherlands. (email: s.yadav@utwente.nl; a.b.j.kokkeler@utwente.nl)

P.A.D. Legaspi, M.S. Oude Alink and B. Nauta are with the Integrated Circuit Design (ICD) group of University of Twente, The Netherlands. (email: p.a.d.legaspi@utwente.nl; m.s.oudealink@utwente.nl; b.nauta@utwente.nl)

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

Digital Object Identifier XX.XX.XX

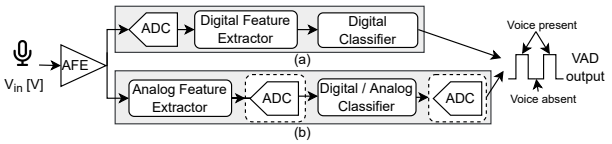

Fig. 2. Block diagram of (a) Digital VAD (b) Analog VAD (presence and positioning of ADC depends on the classifier's implementation.)

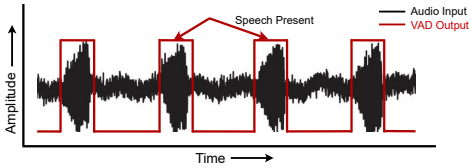

Fig. 3. Output signal from VAD (in red) overlapped on a speech sample. A high level of the VAD output refers to the presence of speech.

sured or simulated at the device level using foundry PDKs) [5]–[25]. This includes digital ([5], [7], [16], [18], [20], [21]), analog ([8], [10]), and mixed-signal ([6], [9], [17], [19], [22]–[25]) designs along with some implementations on Field Programmable Gate Array (FPGA) ([13], [14]), Field Programmable Analog Array (FPAA) ([12]), Digital Signal Processor (DSP) ([15]), and Neuromorphic platforms (e.g. Intel Loihi [26]) ([11]), see Fig. 4. It can be observed that activity in this field is really ramping up. Other VAD designs, e.g. simulated in MATLAB® or Python, have been excluded.

The VAD implementations can be broadly divided into two categories: General Purpose (GP) hardware (includes FPGA, FPAA, DSP and Neuromorphic hardware) and Application Specific Integrated Circuits (ASICs) including dedicated digital, analog and mixed-signal ICs specifically for VAD. The GP hardware provides flexibility in terms of designing, quick prototyping, tuning the design in an iterative manner and amortization of development costs over multiple applications, but provides less room for power and performance optimizations. ASICs, on the other hand, are designed and optimized for their intended purpose (e.g. lower power, smaller area, better performance), but at the price of relatively high development costs, reduced flexibility and longer development trajectories.

None of the existing VAD overview papers [27]–[31] focus on comparing and benchmarking the hardware implementations; instead they focus on different classification techniques [27], [28] or algorithms [29]–[31]. An overview on the design challenges faced by voice-activated IoT devices for the healthcare industry is presented in [32]. This overview addresses the high level requirements and challenges for a VAD in the medical field, but does not provide details about the implementation of such systems. Apart from these VAD-focused papers,

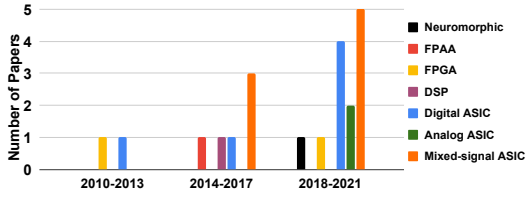

Fig. 4. Literature from past 12 years with different VAD implementations.

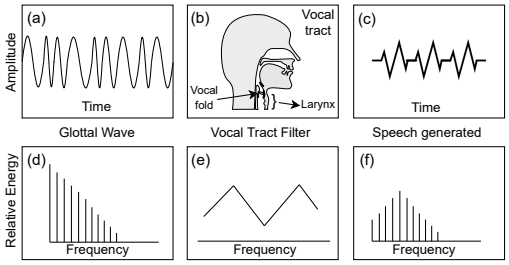

Fig. 5. Source-filter model [37], [38]. (a) Glottal wave (time domain) (b) Illustration of a vocal tract filter (c) Generated speech (d) Glottal wave (frequency domain) (e) Vocal tract transfer function (f) Generated speech.

a tutorial for Sound Event Detection (SED) is presented in [33], which is an extreme generalization of activity detection in the entire human auditory range. A thorough study and detailing of these VAD implementations seems appropriate to better understand different implementation parameters, testing strategies and also to identify the trends and challenges in the field of VAD.

The overview presented here was done by a scoping review process [34], [35]. An exhaustive search conducted in the IEEE and ACM databases resulted in a total of 3575 VAD-related papers. After 2 rounds of title-only and abstract-only screening, 75 papers were selected for the final full-text study. Out of these 75 papers, 21 papers discussed hardware implementations whereas 54 papers discussed the software implementation of VADs.

This overview paper will discuss different designs and testing aspects of VAD along with state-of-the-art hardware implementations. This will highlight the methodology and optimizations used to achieve good classification performance at low power consumption. In addition to this, various datasets and performance metrics used to evaluate VAD will be explained. Lastly, this overview will help readers to understand trends and challenges, and will provide an outlook for expected future developments in VAD designs.

This article is organised as follows: Section II presents the background on human voice production and the definition of common terminology used in VAD designs. Section III presents the literature survey and the metadata obtained from the papers. Section IV provides a thorough discussion on the metadata presented in Section III. Section V concludes this article by providing an outlook on good design and test strategies for future development of VADs.

II. Terminology

A. Human Voice Production

Humans use three different mechanisms to produce voice: respiration, phonation and articulation [36]. Respiration involves bringing air out of the lungs, making them act like a power source. Phonation is the process of air from the lungs flowing through the vocal folds present in the larynx. These vocal folds may open completely to freely pass the air or may (partially) close leading to a vibrating sound known as a glottal wave. Articulation is the process of shaping the glottal wave by the vocal tract, see Fig. 5b, to produce understandable speech.

To better understand these mechanisms, a widely-accepted Source-Filter model [37] is presented in Fig. 5. The Source-Filter model describes the speech production process where the glottal wave shown by (a) and (d) is considered as a source. This glottal wave passes through the vocal-tract filter whose transfer function is depicted by (e). According to this model, understandable speech depicted by (c) and (f) is produced when the source signal propagates through the filter.

The significant portion of the glottal wave is limited within 20Hz - 1200Hz [39], and consists of a low-frequency fundamental tone and its initial harmonics, all of which are used to find the periodicity and pitch of the voice. This glottal wave along with other medium frequency (1kHz - 8kHz) components are used to assist in identifying phonemes, words, etc. [40]. Speech components in the higher frequency region (above 10kHz) are used in source localization, speaker identification and verification applications [41]. The vocal tract filter provides information on the formants, which are defined as the resonant frequencies of different parts of the vocal tract (nasal cavity, oral cavity, etc.). These resonant frequencies are present on the lower side of the speech spectrum along with the glottal wave, and are used to track the vowel height and vowel advancement for ASR [40], [42].

B. Pre-processing and SNR of the speech signal

Speech is a non-stationary signal which has time-varying properties such as amplitude and frequency. Speech present in an environment like air is typically measured using SPL (in dB). A speech signal with a certain SPL is converted to an electrical signal using a microphone which is then passed through the AFE, see Fig. 2. The main purpose of the AFE is to condition the electrical signal by amplifying the low voltage output of the microphone (e.g. the ICS-40310 gives 0.1-112mV$_{rms}$ for 50-112 dB SPL signal at V_{DD} = 0.9-1.3V [43]). SPL observed by the microphone is distance dependent: a drop of 6 dB occurs when distance is doubled [44].

Typically, an AFE contains an LNA with a fixed gain [22]. To accommodate the effect of loudness and distance on SPL, and thus on the microphone's output voltage, a low noise Programmable-Gain Amplifier (PGA) can be used. This programmability in the amplifier can either be achieved by external control signals or by using Automatic Gain Control (AGC) circuitry [6].

In real-life scenarios, as shown in Fig. 6a, the SNR is generally unknown, as input speech (SPL_1) at the microphone is always corrupted by some environmental noises (SPL_2) like traffic, a vacuum cleaner, etc. The SPL value for this corrupted speech signal going into the microphone can be calculated as $SPL_{net} = 10\log_{10}(10^{SPL_1/10}+10^{SPL_2/10}+ ... + 10^{SPL_n/10})$ where SPL values are in dB [45]. The scenario typically considered for testing/evaluating VADs is shown in Fig. 6b: SPL_1 and SPL_2 are separate sound sources, and are considered free from any environmental noise. This scenario allows an SNR value for the total sound signal to be calculated, and is used by tools like Filtering and Noise Adding Tool (FaNT) [46] to combine speech and noise datasets for testing the VADs at different SNR levels. The SNR values calculated here are completely different from the SNR values used for designing the AFEs. For AFEs in speech applications, the sound signal at a particular power level is used as a signal source whereas circuit (with microphone) noise is considered as noise source. The ratio between signal power and circuit noise power is also referred to as "SNR" but should not be confused with SNR as defined in Fig. 6. In this paper, SNR always refers to the former definition as presented in Fig. 6b. Signal-to-noise and distortion ratio (SINAD) is defined as the ratio of signal power to the sum of circuit noise power and distortion power, and is another metric useful for designing analog/mixed-signal circuits. The distortion introduced by the non-linear behavior of physical devices seems somewhat benign for VAD, as it can be compensated by machine-learning or neural-network (NN) based classifiers in the VADs [22]. As SINAD is not (yet) used anywhere, it will not be further discussed.

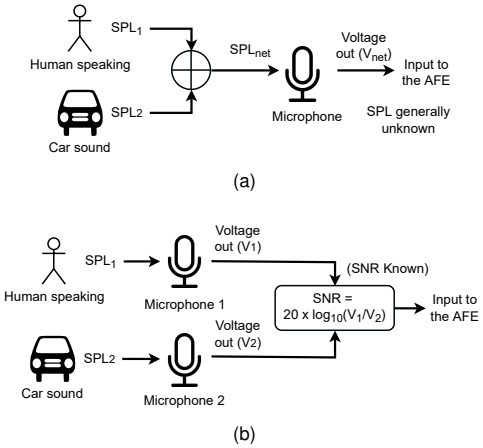

Fig. 6. (a) Addition of different SPL signals before converting them to a voltage signal. (b) Combining the output of different microphones to calculate SNR at the input of AFE.

C. Bandwidth and Dynamic Range

Bandwidth is an important design parameter for analog, mixed-signal as well as digital designs. The sampling rate is generally chosen as the Nyquist frequency required to capture a certain bandwidth. The speech spectrum is considered to lie from 20Hz to 20kHz [47] but the selection of bandwidth or sampling frequency depends on the target application. As discussed earlier, the lower region of the speech spectrum is used to identify the content of speech whereas the higher region is used to identify the speaker characteristics. The presence of speech-related features in the lower and medium frequency region is the primary reason for the widespread usage of 20Hz - 4kHz or 20Hz - 8kHz frequency bands for most of the speech-based applications.

Language is another crucial aspect which is seldom discussed while choosing the bandwidth or sampling frequency. Every language occupies a dominant frequency range towards which the native speakers are very sensitive [48], [49], e.g., UK English has a dominant frequency span of 2 kHz - 12 kHz, whereas Japanese has a non-overlapping span of 125 Hz - 1.5 kHz. The dominant frequency span of different languages is presented in Fig. 7. Considering UK and US English, the

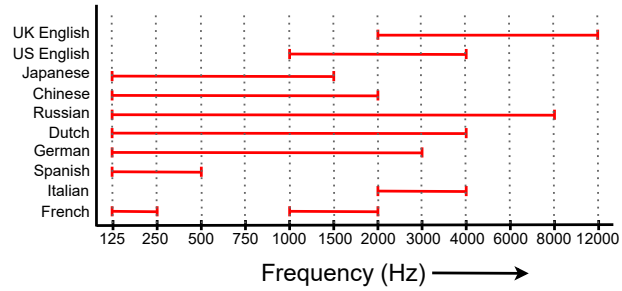

Fig. 7. Range of dominant frequencies occupied by languages [48], [49].

dominant frequency region does not cover the glottal wave region because of difference in phonation and articulation for vowels and consonants. For consonants (21 out of 26 letters in the alphabet), the glottal wave is attenuated which highlights frequency regions other than the glottal wave, see Fig. 7.

The dynamic range (DR) of a system, expressed in dB, is defined as a ratio of the highest signal level a system can handle to the lowest signal level. A healthy human being can perceive sound from the auditory threshold of 0dB SPL [50] to a jet taking off 50m away (140dB SPL) [44] resulting in a DR of 140 dB [44], [51]. However, human ears have a loudness protection mechanism to protect them from hearing loss due to which extreme sound signals cannot be perceived simultaneously, leading to a reduced instantaneous DR. The instantaneous DR for human hearing is estimated to be around 70dB in an average environment but only around 40dB in a very loud environment [51].

From the perspective of VAD systems, the DR of the expected input signal is used to determine the sample width, measured in bits, for the ADC. Based on the 70dB DR of human hearing, a sample width of just over 11 bits is sufficient to digitize the incoming signal. Having a 16 bit resolution to represent the sound signal, as followed by many consumer grade sound standards (for e.g. Compact Discs (CD)), a DR of 98dB can be achieved, which gives plenty of headroom in case of sudden increase in input amplitude.

Sample width has a direct effect on the complexity, size, and power consumption of digital and analog VADs. Apart from DR, desired speech detection accuracy and SNR of the input speech also influence the sample width. Reducing the sample width will put a limit on the maximum accuracy that can be achieved by the VAD systems because reduced bit width will result in higher quantization noise (circuit noise), and thus reduced overall SNR (audio noise + circuit noise).

D. Framing and Overlapping

Speech is a non-stationary signal. The main goal of framing is to make the signal appear stationary in a small time frame enabling the use of the vast signal processing toolbox developed for stationary signals. A too large frame length results in non-stationarity, but also leads to a high latency of the VAD process. A too small frame length will result in low latency VAD operation but will not completely capture the characteristics of speech.

The frames may be overlapped to preserve the temporal information and maintain the rather continuous nature of a speech signal. The amount of computations and the power consumption of VADs are also dependent on the overlap between frames. In simple terms, lower overlap means fewer computations and vice versa. Another term that is used instead of overlapping is “frame shift”. The relationship between overlapping and frame-shift is:

$$\text{Frame Shift (ms)} = \text{Frame Length (ms)} - \text{Overlap (ms)} \quad (1)$$

E. Features

Multiple features used in feature extraction techniques are presented in the literature but all of them can be combined into two broad categories: Amplitude/Energy (A/E) based features and Event-based features. A/E based features involve either some form of amplitude (e.g. raw speech amplitude, amplitude of filter outputs, etc.) or energy (e.g. frame energy, frame energy difference, filter output energies, etc.) of the speech signal. A/E based features continuously track the input speech signal whereas event-based features are asynchronous spikes which are generated when A/E based features are accumulated over time and the accumulation crosses a certain threshold.

Standalone VAD implementations are not very useful but when designed together with applications like KWS or ASR, lower power consumption by these applications can be achieved. KWS and ASR (almost) always make use of a commonly used feature set: the Mel-Frequency Cepstral Coefficients (MFCC). This is a mathematically complex feature set motivated from the source-filter model (see Fig. 5) which is also used for VADs [52], [53]. The usage limitation of features on the FEx architecture will be discussed in Section III-C.

F. Classifiers

We categorize classifiers into 3 types: Decision Rule (DRule), Machine Learning (ML) and NN. DRule-based classifiers compare the output of the FEx to a user-specified or adaptive threshold value to differentiate between speech and non-speech segments.

ML consists of a set of algorithms, such as Logistic Classification, Decision Tree (DT), and Support Vector Machine (SVM), which require a set of distinguishing features explicitly labeled by human experts, and a structured set of data to differentiate between different inputs [54].

NNs are a subset of ML, but here we explicitly distinguish between them to focus on the aspect of human intervention, their learning behavior and data handling capabilities. NNs, such as Deep Neural Networks (DNNs), Binary Neural Networks (BNNs) and Binary Weighted Neural Networks (BWNs), try to mimic the human brain. They can work on both structured and non-structured sets of data, and can also extract distinguishing features on their own without any human intervention during the learning process [54]. NNs require a bigger structured / unstructured set of data to achieve good classification results than other ML algorithms which rely on a structured set to give good classification results.

The classifier’s performance depends a lot on its ability to handle incoming noise. In order to improve the performance, noise-independent (covering all possible noises at different

SNRs) training of the classifiers is performed. It is safe to assume that the noise-independent training will lead to a complex classifier with bigger area and higher power consumption compared to a noise-dependent one. This is exploited by [17] where the BWN is purposely trained for specific noise to save area and power. It is difficult to estimate the power consumption and speech/non-speech detection accuracy degradation when the noise context changes.

G. Speech and Noise datasets

Evaluating a system is as important as designing it. A good validation and testing strategy can help develop better understanding about the design and might open opportunities for future improvements. A validation strategy can help in improving the design by evaluating it and choosing the best performing design for final implementation, whereas a testing strategy will provide performance results for the chosen implementation.

In case of VAD, validation (tuning classifier parameters) and testing (accessing performance of a model) is done by providing either analog (noisy) speech samples to the analog/mixed-signal designs, or digital (noisy) samples to the digital designs. These (noisy) speech samples are either taken from available standard datasets with labeled ground truth or are generated by tools like FaNT from separate standard voice and noise dataset, as discussed earlier in Section II-B.

H. Performance Metrics

To make a fair comparison between VAD designs, we quantify the functional and hardware performance. Functional performance involves the detection accuracy of speech and non-speech segments in an audio signal, whereas hardware performance focuses on the energy, power and latency of the designs. These metrics cannot be trivially combined into a single metric, since there seems to be no fundamental trade-off and the weighing of different factors is application-dependent.

There are some generally accepted metrics used to functionally evaluate VAD designs, which are also used by the articles considered for this review. These metrics are derived from the confusion or error matrix used to visualize the performance of a classification algorithm [55]. Detection Accuracy (DA) is the simplest metric used to analyze the functional performance of VAD but different definitions exist:

$$\mathrm{DA}_1 = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}} \tag{2}$$

$$\mathrm{DA}_2 = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \text{ (Speech detection)} \tag{3}$$

$$\mathrm{DA}_3 = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \text{ (Non-Speech detection)} \tag{4}$$

where TP is the True Positive count, TN is the True Negative count, FN is the False Negative count, and FP is the False Positive count. More information on these terms can be found in [55]. In the case of VAD, where speech is often considered to be sparse in nature, both DA_1 and DA_3 can be maximized for non-speech detection by building a biased classifier which always provides output as "non-speech". The possibility of having biased results hints towards the inability of these metrics to truly capture the classification behavior. Therefore, other metrics are used to quantify the performance.

$$\text{Precision} = \mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{5}$$

$$\text{Recall} = \text{Sensitivity} = \mathrm{TPR} = \mathrm{SHR} = \mathrm{DA}_2 \tag{6}$$

$$\text{F-measure } (\mathrm{F}_\beta) = (1+\beta^2) \times \frac{\text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision}) + \text{Recall}} \tag{7}$$

$$\text{F1-measure (F1-score)} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

Precision or positive prediction value (PPV), as described in eq. (5), quantifies the correct positive predictions made out of total positive predictions, whereas Recall, as described in eq. (6), also known as sensitivity, True Positive Rate (TPR) and Speech Hit Rate (SHR), quantifies the correct positive predictions made out of actual positives, and is equivalent to DA_2. These metrics used together are deemed better than DA_1, DA_2 and DA_3 because with precision, one can make sure that what is identified as positive is actually a positive, whereas recall makes sure that positive observations are not missed. Both precision and recall are important when building a classifier, so, instead of using two separate metrics, the F-measure or F-score [56] combines these metrics, see eq. (7), where β is a positive real value chosen to prioritize either precision or recall based on the application. A balanced F-measure is sometimes also used which is defined as the harmonic mean of precision and recall ($\beta = 1$), see eq. (8). A better classification will result in a higher F-measure.

SHR and Non-Speech Hit Rate (NSHR) are other metrics used together to evaluate the quality of classifiers. NSHR, also known as Specificity or Selectivity or True Negative Rate (TNR) is the detected ratio of non-speech frames and is equivalent to DA_3.

$$\text{NSHR} = \text{Specificity} = \text{Selectivity} = \mathrm{TNR} = \mathrm{DA}_3 \tag{9}$$

Sometimes the performance of a VAD is also shown by a Detection Error Tradeoff (DET) curve and measured in terms of Equal Error Rate (EER). DET is a 2D plot represented by False Positive Rate (FPR) and False Negative Rate (FNR) as the two axes:

$$\mathrm{FPR} = 1 - \mathrm{TNR} = \frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} \tag{10}$$

$$\mathrm{FNR} = 1 - \mathrm{TPR} = \frac{\mathrm{FN}}{\mathrm{TP}+\mathrm{FN}} \tag{11}$$

EER is defined as a point on the DET plot where FPR and FNR are equal. A smaller EER value means the classifier is better. The Half Total Error Rate (HTER), also known as balanced error rate (BER) is similar to EER but instead of equating FPR and FNR in the DET plot, it takes their average:

$$\mathrm{HTER} = \mathrm{BER} = \frac{\mathrm{FPR}+\mathrm{FNR}}{2} \tag{12}$$

Hardware performance metrics generally used for VAD implementations are energy per frame, average power, latency and throughput. Energy per frame for VADs is defined as the energy consumed during the time required to process one frame or to make one speech/non-speech decision. Latency is defined as the time required by the VAD to make a decision from the point where input speech is first provided. Bounds on the latency are decided by the application. According to [23], a frame size of 100ms is good enough to make a real-time VAD. But in an application like hearing aids, latency above 30ms is considered high and annoying for the users [22] whereas a higher latency value (e.g. one second) can be tolerated for a voice activated switch. High latency VADs cannot be used for complex speech processing tasks such as ASR or phoneme detection because the information loss happening due to delayed activation of processing blocks will lead to missed detections. One possible solution is to store the past samples and process them once speech is detected, leading to additional storage requirements.

III. Literature Survey

Table II provides an overview of all the VAD designs covered in this review. In recent years, see Fig. 4, with the emerging demand of ultra-low power VADs for battery-operated portable devices, purely analog designs started appearing in literature along with an increased usage of digital and mixed-signal ASICs. Table I gives a list of implementation domains presented by different papers. A mixed-signal design with an analog FEx and an unknown digital classifier implementation is presented in [9]. A Neuromorphic hardware-friendly Spiking Neural Network (SNN) based VAD is presented in [11] but it is unclear whether actual neuromorphic hardware was used because the published power figures were based on estimates rather than actual measurements. A VAD design for low memory and computation power can be found in [16], but the implementation platform is not explicitly stated by the author. Usage of memory and processor hints either towards a digital or a mixed-signal platform. For this review, we categorize this paper as a digital ASIC design. The VAD design presented by [12] is the only paper in our list using an FPAA. Similarly, [21] is the only paper which made use of an ARM based embedded system to implement a real-time VAD. A mixed-signal ASIC is presented in [22] discussing only the implementation of the analog FEx; details about the classifier were not provided.

Some implementations had no silicon but used the technology PDKs along with software tools to provide the results. An analog FEx in a 180nm process is presented in [23] but the classifier is a MATLAB® based trained SVM model. A mixed-signal design using TSMC's 28nm HPC+ PDK is presented in [24]. An analog FEx with custom instruction based DSP classifier is presented by [25] which uses the Synopsys® HSIM tool to arrive at the simulated results.

Some implementations made use of external components/hardware to make their design functional. An external processor was used by [12] to control the functioning and programming of the FPAA. An analog FEx with mixed-signal DT classifier is discussed in [19] where an ARM Cortex M4 processor is used to generate the VAD output and re-train the DT classifier when the classification accuracy is poor. Due to the usage of additional components to offload the control and processing tasks, the results presented in [19] cannot be considered as a true indicator of the performance of the VAD.

TABLE I
Papers with their respective hardware implementation domains

Category	Implementation Domain	References
GP	Neuromorphic Hardware [a]	[11]
GP	FPAA	[12]
GP	FPGA	[13], [14]
GP	DSP	[15]
ASIC	Analog (or PDK sim.)	[8], [10]
ASIC	Digital (or PDK sim.)	[5], [7], [16], [18], [20], [21]
ASIC	Mixed-signal (or PDK sim.)	[6], [9], [17], [19], [22]–[25]
	Total	**21 papers**

[a] Digital FEx with classifier on Neuromorphic platform

A. Bandwidth and Dynamic Range

Sampling frequency and sample width are important design parameters for a digital VAD but either or both the parameters are not provided in [5], [15], [20]. Mixed-signal designs presented in [9], [19], [22]–[24] do not provide complete information about the bandwidth, sample frequency or sample width.

A few papers mention the use of multiple sampling frequencies and sample widths. Multiple sampling frequencies (from 2kHz to 32kHz) are presented by [16] but which frequency is used to evaluate the VAD is not clearly stated. Multiple sample widths are presented by [24] where an approximate architecture is used to work on the reduced sample width when the SNR of the input signal is high. A mixer-based architecture is used by [6] to sequentially downconvert different bands of audio signals within $\leq$ 4kHz to a passband of 500Hz. The ADC connected after the mixer then runs at 1kHz instead of Nyquist rate of 8kHz to save power.

In mixed-signal designs where FEx is implemented in the analog domain and the classifier in the digital domain, a single/multi-bit ADC is placed at the separation boundary to convert the analog features into digital format. In most of the cases for digital VAD implementations, the ADC is not provided on-chip. Therefore, an external ADC or digitally generated signal is used to provide digital inputs to the VAD. The usage of external components by digital VAD designs makes them hard to directly compare to the analog and mixed-signal VAD designs.

As mentioned earlier in Section II-E, analog and mixed-signal VAD designs mostly make use of amplitude or energy based features which are computed using a bank of filters with different center frequencies and inter-band spacing. Sixteen geometrically spaced analog filters from 100Hz-5kHz are presented by [17]. A claim has been made by [19] along with [58] that most of the energy for speech and acoustic noise is concentrated from 100Hz to 4kHz, thus, making use of 16 exponentially spaced filters with center frequencies from 75Hz to 5 kHz. Computation of the log energy of 128 Mel spaced filter banks is presented by [11].

TABLE II
DESIGN PARAMETERS, FEATURES, CLASSIFIERS AND DATASETS USED BY THE VAD DESIGNS

Paper	Domain	Frame Length (ms)	Overlap (%)	Bandwidth	Sampling Frequency (kHz)	Sample Width (bits)	Features	FEx Category	Classifier	Classifier Category	Voice Dataset	Noise Dataset
[5]	Digital	N/A[a]	N/A	N/A	N/A	–[b]	Signal Energy + Harmonicity[c]+ Modulation Frequency[c]	A/E[d]	Custom NN	NN	AURORA2	AURORA2 Custom[e]
[6]	Mixed-signal	512	0	$\leq$ 4kHz	1	8[f]	Energy of DCT bins	A/E	Custom NN	NN	Libri Speech	NOISEX-92
[7]	Digital	32	68.75	N/A	16	16[g]	Signal Energy + Noise Energy	A/E	Threshold comparator[h]	DRule	Custom	Custom
[8]	Analog	N/A	N/A	300Hz-6.8kHz	–	–	Signal Energy + Noise Energy	A/E	Threshold comparator	DRule	Custom	Custom
[9]	Mixed-signal	20	50	N/A	8	–[i]	Energy difference between neighbouring Mel-shaped filters	A/E	DT	ML	NOIZEUS	NOIZEUS
[10]	Analog	N/A	N/A	300Hz-4kHz	–	–	Filter Bank Amplitude / Energy	A/E	Custom NN	NN	NOIZEUS	NOIZEUS
[11]	Neuromorphic	40	50	N/A	16	–[b]	Log-Mel Filter Bank Energies	A/E	SNN	NN	TIMIT	QUT - NOISE
[12]	Analog	N/A	N/A	100Hz-5kHz	–	–	Filter Bank Amplitude / Energy	A/E	WTA[j]	DRule	TIMIT	Custom
[13]	Digital	40	62.5	N/A	16	16[k]	DoV[c]+ QSNR[c]	A/E[c]	Threshold comparator	DRule	Custom	Custom
[14]	Digital	10	0	N/A	8	16[b]	Frame Energy + Max absolute signal difference + Max absolute signal squares difference	A/E	Logistic Classification	ML	Custom	Custom
[15]	Digital	25	80	N/A	8	24[l]	Filter Bank Energy	A/E	Threshold comparator	DRule	TIMIT	Custom
[16]	Digital	4,8,16,32	75-80	N/A	2 - 32 kHz	16[m]	**No features: Raw speech**	A/E	DNN	NN	Custom	Custom
[17]	Mixed-signal	25	60	100Hz-5kHz	16	16[n]	Signal Energy converted to Events	Event	BNN	NN	AURORA4	DEMAND
[18]	Digital	8	N/A	N/A	8	16[f]	Squared Energy from FFT	A/E	DT	ML	Custom	Custom
[19]	Mixed-signal	N/A	N/A	75Hz-5kHz	N/A	16[o]	Average value of rectified signal in different frequency bands	A/E	DT	ML	NOIZEUS	NOIZEUS
[20]	Digital	8	N/A	N/A	N/A	16[b]	Squared Energy from FFT	A/E	DT	ML	N/A	N/A
[21]	Digital	128	N/A	N/A	8	8[f]	Teager Energy[p] to Energy Ratio	A/E	Threshold comparator	DRule	TIMIT AURORA	NOISEX-92 AURORA
[22]	Mixed-signal	25	60[b]	100Hz-5kHz	N/A	16[n]	Energy computed in analog FEx converted to events	Event	DNN	NN	AURORA4	DEMAND
[23]	Mixed-signal	100	50	100Hz-2kHz	10	8[f]	Energy based feature in analog domain	A/E	SVM	ML	TIMIT	MUSAN
[24]	Mixed-signal	25	60	N/A	N/A	9/8/4/1	Signal Energy converted to events	Event	BWN	NN	TIMIT	NOISEX-92
[25]	Mixed-signal	120	0	125Hz-4kHz	8	8[f]	Spectral Information extracted from different frequency bands	A/E	SVM	ML	Libri Speech	DARPA N-Zero

(a) N/A = Not Available (b) Unknown / Not clearly stated (c) Additional features used to enhance the performance of the design (d) Amplitude or Energy based features (e) Refers to non-standard dataset (f) On-Chip ADC (g) Pulse Code Modulated (PCM) data from external delta-sigma integrator (h) Adaptive Threshold Comparator (i) On-chip fast and slow running ADC; unknown ADC width (j) Winner Takes All (k) External AD1836A codec IC used (l) ADC present on DSP platform (m) Samples taken from .m4a format file (n) No explicit ADC, 16 parallel channels of IAF modules (o) Off-chip ADC and DAC used with unknown width (p) Teager Energy definition presented in [57]

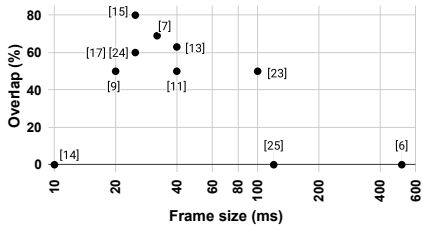

Fig. 8. Combination of different frame lengths and overlapping percentages

B. Framing and Overlapping

Framing and overlapping are the pre-processing techniques applied on the raw speech signal as explained in Section II-D. Fig. 8 presents a scatter plot of frame sizes and overlapping percentages. Only 11 out of the 21 paper give this information. The typical choice seems to be 20ms - 40ms for frame length with 50% - 80% of overlapping, although the number of data points is insufficient to make any hard claims.

The analog design presented in [8] makes indirect use of framing with zero overlap percentage. Frame sizes of 8ms, 16ms, 32ms are used for the analog integrator in FEx to compute the energy of the signal over the selected time frame. The analog design of [22] does not use these aforementioned pre-processing techniques on raw speech but instead uses frames and overlaps the Integrate-And-Fire (IAF) module's output in the FEx to provide it to the digital classifier. This IAF module integrates the incoming energy and generates an asynchronous output when accumulated energy reaches a certain user-specified threshold.

There are different ways of framing and overlapping, some of which can cause information loss potentially leading to a reduction in VAD performance. A frame size of 512ms is used in [6] which is further divided into small non-overlapping sub-frames of 16ms each, leading to a total of 32 sub-frames. Corresponding to every sub-frame, a set of 32 frequencies is generated sequentially and is fed sequentially to a mixer for downconversion. The downconverted signal is then used to calculate energy for every sub-frame. These energy-based features are distributed over time which causes information loss for the design. Similarly, in [25], a frame size of 120ms is used which is divided into 15 non-overlapping sub-frames. Each smaller frame is then used to calculate power at a certain frequency. As only part of the entire frequency information is covered for a given time frame, a lot of useful information in the speech spectrum gets lost.

In order to achieve low power and real-time performance, VAD systems require (jointly) optimized values for frame size and overlap percentage. Bigger frame sizes of 120ms and 512ms in [6] and [25] lead to a non real-time design. A moderate latency of 100ms proposed by [23] is only acceptable for a subset of applications (e.g. voice operated switch) when speech is sparse in nature. An upper bound of 30ms can be placed on the frame size and latency as per hearing-aid user's experience [22].

C. Features

A limited set of features is implemented by analog and mixed-signal VADs. Most FEx architectures use a bank of

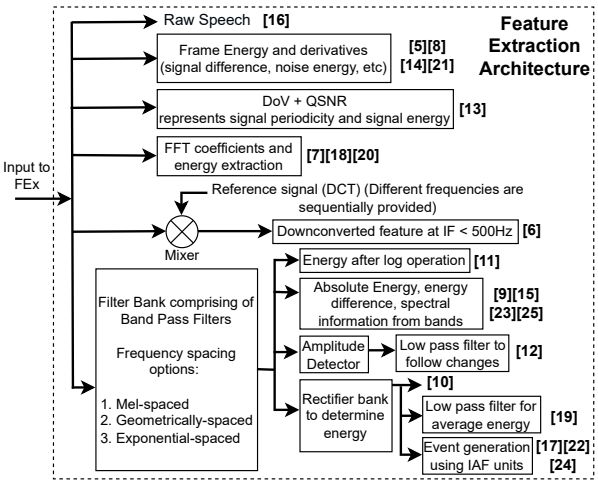

Fig. 9. Feature Extraction Architecture implemented by VAD designs

geometrically or exponentially spaced Band-Pass filters (BPFs) to extract information in different frequency bands, see Fig. 9. In A/E based systems, these BPFs are followed by a rectifier circuit to extract the energy in each frequency band. For an event-based FEx, an IAF module is placed after the filter banks to convert the amplitude/energy accumulation to asynchronous events, which are then used by a digital classifier, thus making it a mixed-signal design [17], [22], [24].

In some cases, additional features are used when the speech signal gets degraded in noisy environments. A design presented by [5] uses 2 additional features (Harmonicity and Modulation Frequency) apart from signal energy as the main features. Degree of Voicing (DoV) [59] along with Quantile-SNR (QSNR) [60] are used by [13] as features of the VAD. DoV represents periodicity of a speech signal and is considered as a good representation for speech but it suffers from periodic noise. To avoid this, [13] makes use of QSNR to compensate for the performance degradation.

A claim has been made by [15] along with [61] stating that under heavy noise situations (0dB SNR), the first formant lying in the lower frequency region can still be detected and used for VAD. According to [19], the middle frequency features are good enough to do VAD; low and high frequency features can be used in the low SNR range or for other tasks such as ASR and SV. Lower and upper bounds for these frequency regions cannot be clearly established from the literature. According to the experiments conducted by [19], increasing the number of filters beyond 16 does not improve the classification accuracy of the VAD. A mixer-based FEx is presented by [25] which downconverts the features present at higher frequencies to a lower frequency band. A band-limited filter is later used to extract features. This approach seems to save power for the VAD operation.

A reconfigurable time-multiplexed BPF is presented in [23] to process sub-frames of 10ms each for different frequency regions over a frame size of 100ms. As this extraction is 10% of the total speech information, it works for higher SNR inputs but might fail for low SNR inputs.

An attempt to avoid calculating features representing redundant information about speech for VAD tasks has been

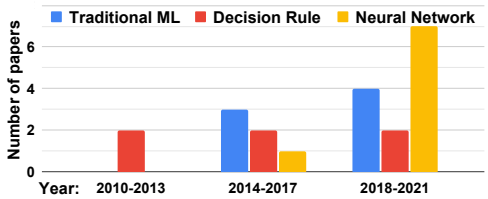

Fig. 10. Trend in the usage of different classifiers for VAD

reported by [9] and [19]. Based on the current noise context, mutually exclusive features are used to separate speech and non-speech segments. Care should be taken in such VAD techniques to ensure that the power consumption to detect context-change and re-configuring the VAD does not offset the benefits.

D. Classifiers

The usage of different classifiers over the years is presented in Fig. 10, showing increased popularity of NN-based classifiers in recent years. This can be attributed to the fact that they can achieve good accuracy with both structured and unstructured data without human intervention. Usage of NNs also opens up the possibility to remove the FEx stage, as presented by [16]. A NN can be trained on raw speech signals which provides it with the flexibility of choosing its own features to produce desired results. However, this flexibility will come at the expense of increased power consumption due to additional hidden layers required in the NN for feature extraction.

Implementation of NNs differ from paper to paper. A flexible NN classifier is implemented by [6] using a custom digital processor. This processor was designed with a custom instruction set and specialized hardware for Matrix-Vector Multiplication (MVM), FFT, element-wise vector multiplication, non-linear activation, and min/max/avg operations. Based on the available memory, an arbitrary NN model can be supported by this processor. A generic DNN is used by [22] to test the VAD without giving any implementation details. An off-chip Matlab® based pre-trained SVM classifier is used by [23] to produce the VAD outputs. A custom-built Fully-Connected (FC) BWN is presented by [24] where the bit-width of the BWN classifier is adjusted based on the SNR prediction module which operates on the incoming speech. Dynamically changing the bit width based on the correctly predicted SNR value can lead to power savings in high SNR conditions.

A threshold comparator is the simplest piece of hardware which can work like a (binary) classifier and can have a very low power consumption. A dual threshold comparator is presented by [15] to perform the classification task. Dual thresholds are intended to compare lower and higher frequency regions of speech separately to get better VAD results. A nearly-fixed threshold comparator is presented by [21]. The features are shown to be robust to stationary noise which does not require much changes in the threshold of the comparator but changes in threshold values for non-stationary noise are not discussed in [21].

DT classifiers are simpler than NNs and are trained with a labeled dataset where an algorithm called “C4.5” is generally used [62]. DT classifiers are presented as a good choice for hardware implementation by [9] because the training algorithm can be modified (modification of cost-function and pruning) to account for context-aware and dynamic resource-cost-aware feature selection through machine learning methods. This will enable the system to activate only those features which are relevant for a given context, leading to lower power consumption. Similarly, [19] also modifies the training algorithm to maximize the information-gain per Watt resulting in a resource-efficient model. These modifications help to achieve better DT performance with less hardware and lower power consumption.

DT classifiers also have a lot of downsides. DT classifiers making decisions on a single speech frame can lead to a high output toggle rate for low-SNR speech signals which necessitates the usage of a post-processing mechanism to smooth out the VAD decision [10], [19]. This mechanism can be implemented in multiple ways, e.g. to look at the VAD decisions over past and future frames (also known as hangover time) to get a better result [9], [13], [17], [25] or make use of a low-pass filter (LPF) with very small bandwidth to get a smooth VAD output [10]. Both of these approaches result in additional latency.

In [20], it is stated that a large DT improves accuracy at high SNR but due to over-fitting at lower SNR, the performance drops. If the DT is pruned to reduce its size, the maximum achievable performance will also reduce due to under-fitting. In order to compensate for the lower performance, [20] uses frame energy as an additional feature along with a post-processing mechanism.

As mentioned earlier in Section II-F, noise-independent training can lead to higher power and larger area requirements. Care should be taken to design VADs which can adapt to all the relevant noise scenarios. Noise-dependent training of classifiers will lead to a VAD design which probably will not perform well in real-life scenarios.

E. Speech and Noise datasets

Various standard datasets and custom-recorded audio samples are used to evaluate the performance of VAD, see Table II. The most commonly mentioned standard dataset for speech is TIMIT [63], and for noise is NOIZEUS [64]. NOIZEUS is not a pure noise dataset but uses the AURORA database [65] to create noisy speech samples.

Most VAD implementations make use of custom-recorded speech samples and noise samples for evaluation. For example, loud ambient noises like a car, a fan and a busy road from an unknown source are used by [8] instead of standard noise datasets. Similarly, [7], [18] make use of a custom recorded and manually labelled dataset for their VAD evaluation. The VAD design presented in [13] makes use of only one sentence for training and evaluation which could mean that the performance may significantly degrade in real-life scenarios. The credibility of these datasets cannot be established because the information about the demographics and recording environment is either missing or incomplete. In contrast to this, the VAD design presented by [16] provides

TABLE III
DIFFERENT PERFORMANCE METRICS USED BY VAD DESIGNS

Performance Metrics [a]	References	#Papers
SHR vs. NSHR	[9], [10][b] [6], [17]–[20], [22]–[24]	10
DA_1 / DA_2 / DA_3	[8], [9], [12], [16], [21], [25]	6
F-measure	[15], [18]	2
EER	[5]	1
HTER	[11]	1
DET	[5], [7], [8]	3
Figure showing overlap between speech and VAD output	[9], [10], [14]–[17], [21], [23]	8

[a] [13], [14] does not give any quantitative performance metric
[b] [9] gives only plot for SHR vs. NSHR

recording and processing information along with the details about the speakers.

A few VAD implementations make partial use of standard datasets for evaluation, e.g., a single sentence from TIMIT with Additive White Gaussian Noise (AWGN) was used by [12], and a subset of TIMIT and DEMAND was used by [15] and [17] respectively. On the other hand, [20] does not mention anything about the dataset. Usage of a custom dataset or partial usage of a standard dataset should not be considered as a standard practice for evaluating the VADs. A lack of standard testing guidelines for a data-dependent application like VAD leads to performance and power figures for various implementations that cannot be compared.

Another important aspect involved while evaluating the VAD designs is the SNR of the noisy input speech signal. Indicating the input signal's SNR during VAD evaluation is important because it shows the noise immunity of the system. To make a robust VAD design, it is important to evaluate the VAD in low as well as high SNR conditions. In the speech processing domain, speech signals having $\geq$ 20dB SNR are considered as clean speech samples which puts an upper limit on the SNR for VAD evaluation [12], [66]. A lower limit for the SNR does not exist to evaluate the VADs but based on the reviewed papers, it can be set to -5dB [5].

F. Performance metrics

The evaluation of VAD is quantified by the metrics mentioned in Section II-H. The results are presented in Table III. Conversion between these metrics is not possible, making it difficult to compare designs. Some papers present time-domain plots showing the overlap between speech and VAD output but without any quantitative entity. This also makes the comparison of VAD performance impossible.

SHR vs. NSHR and DA are two prominent metrics used for the evaluation of the VAD. Other metrics like F-measure, EER, HTER and DET are not used very often. A plot between NSHR vs. SHR for different VAD designs at different SNR values is given in Fig. 11. Multiple definitions of DA exist (see Section II-H), but is often not explicitly stated by the authors. So, for simplicity, all the DA terms are treated identically. Table IV provides DA values for different VAD designs. The

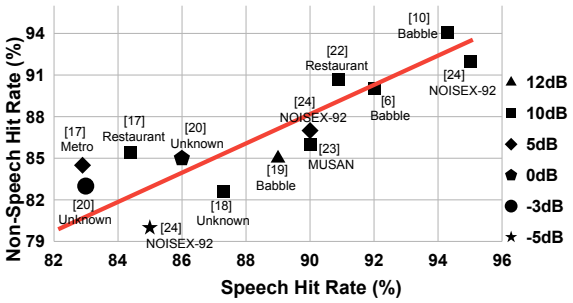

Fig. 11. NSHR vs. SHR at different speech signals corrupted by various noise sources at given SNR values, and trendline indicating classifiers biased towards SHR

TABLE IV
ACCURACY OF DIFFERENT VAD IMPLEMENTATIONS

Reference	Accuracy (%)	SNR Level (dB)
[8]	99.5	N/A
[9]	72 - 88	5
[12]	99.9 / >70 / $\sim$ 2	20 / >7 / <7
[16]	71.7	N/A
[21]	97.8 / 97.1 / 89.4	5 / 0 / -5
[25]	$\sim$79.0 / $\sim$86.5 / $\sim$91.5 / $\sim$93.5	0 / 2 / 4 / 6
[25]	$\sim$94.0 / $\sim$95.5 / $\sim$96.5 / $\sim$97.0	8 / 10 / 12 / 14
[25]	$\sim$97.3 / $\sim$97.3 / $\sim$97.5	16 / 18 / 20

$\sim$ Estimated value from the plots

high DA values of 99.5% and 99.94% by [8] and [12] indicate a biased classifier towards speech detection at high SNR, as the DA value drops to <2% when SNR is still at a relatively high value of $\sim$7dB. The SNR value is not stated in [8].

A plot highlighting the power consumption of VAD designs against the average of SNR and NSHR values (from Fig. 11) is presented in Fig 12. The designs presented by [10] and [24] have the best performance, but also highest power consumption. The next best candidates in terms of average SHR and NSHR value are [6] and [22], which actually have the lowest power consumption. More power consumption is expected to achieve better performance, but these contrasting results makes it difficult to establish such a trend between the power consumption and widely used performance metric of SHR vs. NSHR. Additionally, no trend is visible for different technology nodes in Fig 12.

A plot between power and latency is provided in Fig. 13. There were a few papers which mention only the latency and cannot be included in Fig. 13. A latency of 800ms is

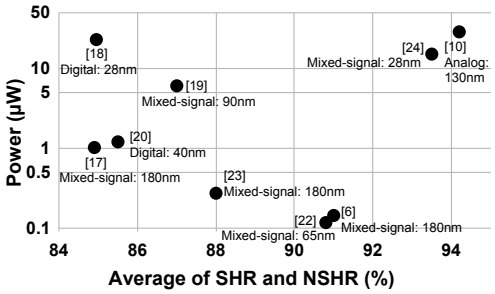

Fig. 12. Average of SHR and NSHR plotted against power consumption for a given technology node

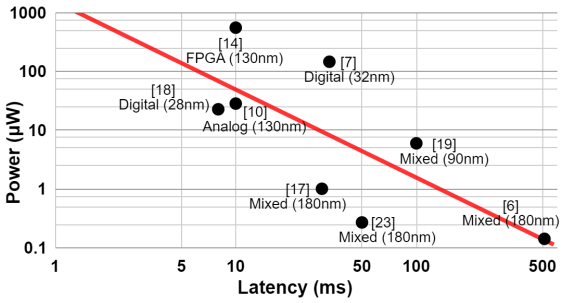

Fig. 13. Power vs. Latency Plot for VAD designs in different implementation domain and technology (Red line highlights the trend)

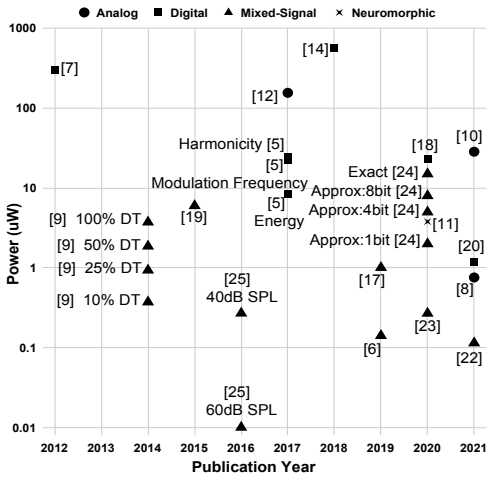

Fig. 14. Power consumption of VAD designs over the years (different versions of same design also presented)

given by [13], an upper bound on latency of 40ms is given by [15], variable latency numbers between 100ms and 500ms are presented by [5]. Latency values are not mentioned by [16] and [21]. Instead of latency, sometimes throughput is also given [24], [25]. To properly analyze the real-time performance of VAD, latency is a more important parameter than throughput. Even with the limited number of data points, there is clearly a trade off between these entities: a lower power consumption can be achieved if higher latency can be accommodated.

To maintain a certain accuracy at lower SNR, [20] uses additional features which leads to an increased latency of 88ms at -3dB SNR from $\leq$ 56ms at $\geq$ 0dB SNR. Usage of additional features (extra hardware) leads to more power consumption but [20] only mentions power of 1.19μW for $\geq$ 0dB SNR. Only the power of FEx is provided by [23], the power figure for the classifier is unknown.

The power consumption of different VAD implementations over the years is plotted in Fig.14. The power figures mentioned by [9] only consider the power of FEx. The active state of the DT classifier mentioned for [9] in the figure refers to the power consumed by the FEx when the classifier is trained multiple times for change in maximum permissible power for FEx. So, 100% DT means the classifier is trained for FEx with default power consumption, 50% DT means the classifier is trained for FEx running with only half the power, and so on. This reduction in power for FEx is achieved by focusing only on the discriminative features to maximize the information gain per Watt. The VAD design by [5] provides power numbers for different features used for classification. A concept of exact and approximate computing is presented by [24] where the power consumption reduces when the design uses approximate computing with 1/4/8 bits in contrast to the exact computing with 9 bits.

The power numbers presented by the mixed-signal design of [25] are based on the simulations done using TSMC 180nm technology for input signals with 40dB SPL and 60dB SPL. Similarly, an estimation of power numbers is given by [11] for the SNN classifier on Neuromorphic hardware. Without knowing the simulation parameters and assumptions made during the estimation process, the power figures presented cannot be used for comparison purposes.

The usage of external hardware is mentioned by [12], [19]. An MSP430 microcontroller is used by [12] to make the final VAD decision based on the intermediate speech and noise output of the WTA algorithm without any mention of power consumption or division of power across different components. A Cortex M4 microcontroller is used by [19] to detect the context change in the incoming signal, retrain the classifier to accommodate the change and generate the final VAD output. These essential VAD computations are offloaded to an external processor with an estimated power of around 57μW. This usage of external hardware for offloading the computations shows that the power figures presented by these VAD designs can not be easily considered for comparison purposes.

In any case, there does not seem to be a trend which highlights reduced power consumption of VADs over the years. With advancement in technology nodes and newer algorithms for VAD, absence of such a trend is astonishing.

IV. Discussion

A. Inputs

A microphone is used in real-life scenarios, whereas a DAC, FPGA or arbitrary waveform generator is used in laboratories to provide an input to the VAD. The possibility to connect a microphone to the VAD is acknowledged by multiple authors. However, details of the microphone and corresponding results are not presented in the papers.

The focus of all the VAD implementations is to either improve the functional performance or the power figures, but a change of perspective is required to understand these performance figures. State-of-the-art commercially available low power microphones [43], [67] are active MEMS based microphones with an average power consumption of 14.4 - 136μW. The lowest-power VAD consumes 10.1 nW [25], three orders of magnitude lower than the lowest-power microphone available in the market. The power consumption of microphones completely overshadows any form of optimizations made in VAD. One possibility is to tightly integrate the VAD with the microphone, as also presented in [67].

A phenomenon called Front End Clipping (FEC) is observed when VAD is accompanied with more complex speech processing applications like a phoneme detector, KWS or ASR. This clipping of initial frames of an audio signal happens due to the time taken by VAD to detect speech and activate

complex components of the speech application. The higher the latency of VAD, the more FEC will be observed. This FEC can lead to missed phoneme or word detection. Appropriate buffering arrangements should be made as per the target application to avoid FEC. This issue is explicitly stated and taken care of by the VAD design presented in [6].

B. Architectures and Implementation Techniques

Authors of different papers have highlighted different techniques to save power. A (passive) switched capacitor implementation is suggested by [8], [10] which is an interesting technique to be considered further for the analog or mixed-signal implementations of VAD. Switched-capacitor-based BPFs are used by [10] to extract information from different frequency bands, whereas [8] uses this technique to perform noise level update and energy scaling in their VAD system.

The division of VAD into several stages and sequential hierarchical activation of stages from low-complexity to high-complexity is claimed to save a lot of power [19]. Another interesting concept of Energy-Quality (EQ) scaling has been used by [18] where the system's energy consumption is traded off with the quality of results required at the output. Different parameters like bias current, ADC resolution, voltage scaling and the number of active nodes in the classifier can be tuned to save energy while still achieving a minimum performance figure required by the application.

According to [17], analog designs suffer a lot from non-linearity and mismatch issues. The biological cochlea is a highly non-linear system, which means that any analog system built to mimic it or to extract some features could potentially take advantage of the non-linearity aspect of the analog designs. This is unlike traditional analog design where linearity is of high importance and requires a lot of design efforts. A VAD design based on the non-linear cochlea open up a lot of possibilities to reduce design effort and power consumption. It is shown by the author of [17] that the non-linear behavior of analog systems can actually be exploited without compromising any performance of the system. Other factors like noise and mismatch still need to be addressed, which may necessitate the usage of compensation and equalization techniques. These issues also have a huge impact on the classifier's performance. Compensation, trimming and equalization techniques, and/or PVT-robust analog design is needed to avoid chip-specific training of classifiers which is an impractical approach for mass-produced designs.

C. Testing

In order to perform a fair assessment of performance and power between the data-dependent application like VAD itself, the primary requirement is to standardize the datasets and testing conditions. A good and thorough testing strategy is currently missing to properly evaluate the VADs. Apart from achieving the best functional performance, focus should also be placed on testing the electrical properties of the system. A good attempt has been made by [8] to test the VAD design with supply noise apart from signal noise. Another attempt has been made by [6] to do an acoustic test of VAD in a sound chamber which tries to mimic the performance of a VAD system in real-life situations.

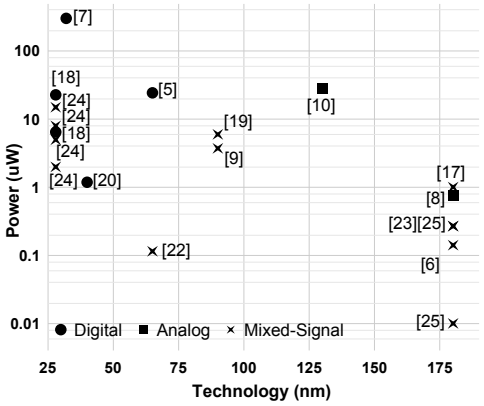

Fig. 15. Increasing power consumption against shrinking of technology nodes for VAD designs

Multiple performance metrics have been used to evaluate the VAD but a standard metric or Figure-of-Merit (FoM) is not available. A FoM seems impossible due to the lack of fundamental trade-offs, but some new metrics combining functional and hardware performance can perhaps be thought of. Most importantly, the datasets and testing conditions need to be standardised.

D. Miscellaneous

Process technology also plays a vital role in designing the circuits. As the technology shrinks and supply voltage reduces, digital designs keep getting better while the analog designs do not scale very well [19]. In contrast to this, technology nodes with bigger feature size typically has lower leakage power and is well suited for analog and mixed-signal designs, as seen in Fig.15. These points should be considered while designing and comparing digital, analog and mixed-signal designs.

V. Conclusion

Voice Activity Detection is an essential function in low power speech recognition applications. Depending on the application (microphone power excluded), the always-on VAD can be the power-dominating component. A good VAD design should have low power consumption to enable its usage in tight power budget applications. It should have low latency to be used in a real-time speech application. A set of 21 VAD hardware implementations within the last 12 years (2010-2021) were discussed. Most implementations are digital or mixed-signal ASICs targeted to reduce power consumption, and most of them made use of Amplitude/Energy-based features along with a neural network based classifier to achieve good performance. A direct trade-off between latency and power consumption is observed. However, the usage of different standard and custom datasets along with varied performance metrics makes it impossible to properly compare or benchmark these VAD designs.

A high SHR and NSHR value (ideal is 100% for all SNR values) should be achieved to have better performance than

TABLE V
RULES FOR UPCOMING VAD DESIGNS

Rule / Parameter	Required/ Values	Remark
Microphone	✓	Testing and reporting performance and power numbers with and w/o microphone.
Voice Dataset	TIMIT	Non-standard / Custom recorded samples are not encouraged.
Noise Dataset	AURORA (stationary + non-stationary)	Non-standard / Custom recorded samples are not encouraged.
Dataset coverage	Full	Partial usage of datasets is not allowed.
Audio SNR	-5dB to +20dB	In steps of 5dB, generated using FaNT or similar tool.
Performance metric	SHR vs. NSHR	If other metric is used, confusion matrix should be provided for conversion to SHR and NSHR.

TABLE VI
GUIDELINES FOR UPCOMING VAD DESIGNS

Rule / Parameter	Values	Remark
Frame size	$<$ 30ms	Frame size + Processing time $\leq$30ms (Real-time design requirement).
Overlap percentage	40% - 80%	Not a strict rule, can be adjusted to give better results.
Sample Width	$\leq$ 12 bits	70dB DR of human ears.
Sampling Frequency	$\leq$ 16 kHz	Based on English language, modify the value for other languages.
Implementation domain	ASIC/GP	ASICs to compare power and latency. GP only for functional correctness. Estimations are not encouraged.
Target Latency	$\leq$ 30ms	Real-time design requirement for any speech application (VAD / KWS / ASR).
Target (Current upper limit) SHR vs. NSHR	94.3% / 94.1% @ 10dB SNR	Both values should be within 1% in order to avoid biased classifiers.
Target power (Microphone power excluded)	$\leq$ 1.01 μW	Based on Fig. 13 for real-time performance of VAD only. If another application is chosen, target power will change.

state-of-the-art. While developing a VAD device, the power consumption of a microphone (current lowest of 14.4μW) should also be considered to evaluate the real life performance of the VAD.

To compare the performance of VAD designs, standardization is required. A standard speech dataset like TIMIT and a noise dataset like AURORA should be used completely and results for all the SNR points should be evaluated and reported in the paper. The use of custom datasets should be discouraged. Non-idealities and mismatches in the analog designs along with the effect of supply noise on the system's performance should also be addressed properly.

The future development of (real-time) VAD designs should adhere to some rules and guidelines for designing and testing of these designs. Table V and Table VI provides a proposed list of rules, parameters and their state-of-the-art values established based on the trends and numbers obtained from this review. If they are closely adhered to by the upcoming VAD designs, a common scale for comparison can be established.

ACKNOWLEDGMENTS

The authors would like to thank Robbert van der Wal from Syntiant Corp for fruitful initial discussions.

REFERENCES

[1] M. Price *et al.*, "14.4 A scalable speech recognizer with deep-neural-network acoustic models and voice-activated power gating," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 244–245.
[2] S. Zheng *et al.*, "An Ultra-Low Power Binarized Convolutional Neural Network-Based Speech Recognition Processor With On-Chip Self-Learning," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 12, pp. 4648–4661, 2019.
[3] J. S. P. Giraldo *et al.*, "Vocell: A 65-nm Speech-Triggered Wake-Up SoC for 10-μW Keyword Spotting and Speaker Verification," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 4, pp. 868–878, 2020.
[4] R. Venkatesha Prasad *et al.*, "Comparison of voice activity detection algorithms for VoIP," in *Proceedings ISCC 2002 Seventh International Symposium on Computers and Communications*, 2002, pp. 530–535.
[5] M. Price *et al.*, "A Low-Power Speech Recognizer and Voice Activity Detector Using Deep Neural Networks," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 1, pp. 66–75, 2018.
[6] S. Oh *et al.*, "An Acoustic Signal Processing Chip With 142-nW Voice Activity Detection Using Mixer-Based Sequential Frequency Scanning and Neural Network Classification," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 3005–3016, 2019.
[7] A. Raychowdhury *et al.*, "A 2.3nJ/frame Voice Activity Detector based audio front-end for context-aware System-on-Chip applications in 32nm CMOS," in *Proceedings of the IEEE 2012 Custom Integrated Circuits Conference*, 2012, pp. 1–4.
[8] M. Croce *et al.*, "A 760-nW, 180-nm CMOS Fully Analog Voice Activity Detection System for Domestic Environment," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 3, pp. 778–787, 2021.
[9] S. Lauwereins *et al.*, "Ultra-low-power voice-activity-detector through context- and resource-cost-aware feature selection in decision trees," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6.
[10] U. Mukherjee *et al.*, "A 28.5µW All-Analog Voice-Activity Detector," in *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2021, pp. 1–5.
[11] G. Dellaferrera *et al.*, "A Bin Encoding Training of a Spiking Neural Network Based Voice Activity Detection," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 3207–3211.
[12] S. Shah *et al.*, "Low power speech detector on a FPAA," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017, pp. 1–4.
[13] J. Jung *et al.*, "A voice activity detection system based on FPGA," in *ICCAS 2010*, 2010, pp. 2304–2308.
[14] G. Meoni *et al.*, "A low power Voice Activity Detector for portable applications," in *2018 14th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME)*, 2018, pp. 41–44.
[15] N. Lezzoum *et al.*, "Voice activity detection system for smart earphones," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 4, pp. 737–744, 2014.
[16] T. W. Sen, "Voice Activity Detector for Device with Small Processor and Memory," in *2019 International Conference on Sustainable Engineering and Creative Computing (ICSECC)*, 2019, pp. 212–217.
[17] M. Yang *et al.*, "Design of an Always-On Deep Neural Network-Based 1-μW Voice Activity Detector Aided With a Customized Software Model for Analog Feature Extraction," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1764–1777, 2019.
[18] J. H. Teo *et al.*, "Low-Energy Voice Activity Detection via Energy-Quality Scaling From Data Conversion to Machine Learning," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 4, pp. 1378–1388, 2020.
[19] K. M. H. Badami *et al.*, "A 90 nm CMOS, 6 μW Power-Proportional Acoustic Sensing Frontend for Voice Activity Detection," *IEEE Journal of Solid-State Circuits*, vol. 51, no. 1, pp. 291–302, 2016.
[20] J. H. Teo *et al.*, "Voice Activity Detection with >83% Accuracy under

SNR down to -3dB at 1.19μW and 0.07mm^2 in 40nm," in *2020 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2020, pp. 1–3.
[21] M. Hadi *et al.*, "An Efficient Real-time Voice Activity Detection Algorithm using Teager Energy to Energy Ratio," in *2019 27th Iranian Conference on Electrical Engineering (ICEE)*, 2019, pp. 1420–1424.
[22] M. Yang *et al.*, "Nanowatt Acoustic Inference Sensing Exploiting Nonlinear Analog Feature Extraction," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 10, pp. 3123–3133, 2021.
[23] E. Shi *et al.*, "A 270 nW Switched-Capacitor Acoustic Feature Extractor for Always-On Voice Activity Detection," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 3, pp. 1045–1054, 2021.
[24] B. Liu *et al.*, "A Background Noise Self-Adaptive VAD Using SNR Prediction Based Precision Dynamic Reconfigurable Approximate Computing," in *Proceedings of the 2020 on Great Lakes Symposium on VLSI*, ser. GLSVLSI '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 271–275. [Online]. Available: https://doi.org/10.1145/3386263.3407589
[25] Y. Chen *et al.*, "A Dual-Stage, Ultra-Low-Power Acoustic Event Detection System," in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, 2016, pp. 213–218.
[26] M. Davies *et al.*, "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, 2018.
[27] S. Nivetha, "A Survey on Speech Feature Extraction and Classification Techniques," in *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, pp. 48–53.
[28] A. Ivry *et al.*, "Evaluation of Deep-Learning-Based Voice Activity Detectors and Room Impulse Response Models in Reverberant Environments," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 406–410.
[29] X. Yang *et al.*, "Comparative Study on Voice Activity Detection Algorithm," in *2010 International Conference on Electrical and Control Engineering*, 2010, pp. 599–602.
[30] Z. Hao *et al.*, "Research of voice activity detection algorithm," in *2011 International Conference on Computational and Information Sciences*, 2011, pp. 853–855.
[31] B. Y.K *et al.*, "Development of robust VAD schemes for Voice Operated Switch application in aircrafts: Comparison of real-time VAD schemes which are based on Linear Energy-based Detector, Fuzzy Logic and Artificial Neural Networks," in *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, 2016, pp. 191–195.
[32] P. Spachos *et al.*, "Voice activated IoT devices for healthcare: Design challenges and emerging applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 7, pp. 3101–3107, 2022.
[33] A. Mesaros *et al.*, "Sound Event Detection: A Tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.
[34] "Systematic and systematic-like reviews," (Accessed on: 21-11-2022). [Online]. Available: https://libguides.csu.edu.au/systematicreviews/scoping
[35] "Prisma for scoping reviews," (Accessed on: 21-11-2022). [Online]. Available: http://www.prisma-statement.org/Extensions/ScopingReviews
[36] C. Anderson, "2.1 How humans produce speech," Mar 2018, (Accessed on: 21-11-2022). [Online]. Available: https://essentialsoflinguistics.pressbooks.com/chapter/2-2-how-humans-produce-speech/#:~:text=Speech%20is%20produced%20by%20bringing,mouth%20and%20nose%20(articulation).
[37] C. Solomon *et al.*, "Objective methods for reliable detection of concealed depression," *Frontiers in ICT*, vol. 2, 04 2015.
[38] Wikimedia, "File:vocaltract.svg," (Accessed on: 21-11-2022). [Online]. Available: https://commons.wikimedia.org/wiki/File:VocalTract.svg
[39] R. B. Monsen *et al.*, "Study of variations in the male and female glottal wave," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 981–993, Oct. 1977. [Online]. Available: https://doi.org/10.1121/1.381593
[40] J. Schnupp *et al.*, *Auditory Neuroscience: Making Sense of Sound*. The MIT Press, 11 2010. [Online]. Available: https://doi.org/10.7551/mitpress/7942.001.0001
[41] B. B. Monson *et al.*, "The perceptual significance of high-frequency energy in the human voice," *Frontiers in Psychology*, vol. 5, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2014.00587
[42] "Lecture 1-7: Source-filter model - University College London," (Accessed on: 21-11-2022). [Online]. Available: https://www.phon.ucl.ac.uk/courses/spsci/acoustics/week1-7.pdf
[43] InvenSense, "ICS-40310," (Accessed on: 21-11-2022). [Online]. Available: https://invensense.tdk.com/wp-content/uploads/2015/02/ICS-40310-datasheet-v1.2.pdf
[44] E. Sengpiel, "Decibel Table - SPL - Loudness Comparison Chart - Sengpielaudio Sengpiel Berlin," 2019, (Accessed on: 21-11-2022). [Online]. Available: http://www.sengpielaudio.com/TableOfSoundPressureLevels.htm
[45] E. Sengpiel, "Adding acoustic levels of sound sources - Sengpielaudio Sengpiel Berlin," 2019, (Accessed on: 21-11-2022). [Online]. Available: http://www.sengpielaudio.com/calculator-spl.htm
[46] J. Kim *et al.*, "Voice activity detection based on multi-dilated convolutional neural network," in *ICMSCE 2018*. New York, NY, USA: Association for Computing Machinery, 2018, p. 98–102. [Online]. Available: https://doi.org/10.1145/3185066.3185086
[47] Digi-Key, "Analog Devices Inc. ADMP803JCEZ-RL," (Accessed on: 21-11-2022). [Online]. Available: https://www.digikey.com/en/products/detail/analog-devices-inc/ADMP803JCEZ-RL/4377078
[48] "Learning foreign languages more easily," (Accessed on: 21-11-2022). [Online]. Available: https://atlantis-vzw.com/foreign-languages.html
[49] SpeedLingua, "Oral skills," (Accessed on: 21-11-2022). [Online]. Available: http://home.speedlingua.com/en/oral-skills/
[50] W. Hamby, "Ultimate sound pressure level decibel table," (Accessed on: 21-11-2022). [Online]. Available: http://www.makeitlouder.com/Decibel%20Level%20Chart.txt
[51] R. Triggs, "What you think you know about bit-depth is probably wrong," 2021, (Accessed on: 21-11-2022). [Online]. Available: https://www.soundguys.com/audio-bit-depth-explained-23706/
[52] R. G. Borin *et al.*, "Voice activity detection using discriminative restricted Boltzmann machines," in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 523–527.
[53] Y. Bai *et al.*, "Voice activity detection based on time-delay neural networks," in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1173–1178.
[54] E. Kavlakoglu, "AI vs. machine learning vs. deep learning vs. neural networks: What's the difference?" May 2020, (Accessed on: 21-11-2022). [Online]. Available: https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks
[55] K. Markham, "Simple guide to confusion matrix terminology," Feb 2020, (Accessed on: 21-11-2022). [Online]. Available: https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
[56] T. Wood, "F-score," May 2019, (Accessed on: 21-11-2022). [Online]. Available: https://deepai.org/machine-learning-glossary-and-terms/f-score
[57] E. Kvedalen, "Signal processing using the teager energy operator and other nonlinear operators," 2003.
[58] F. J. Fahy, "Measurement of acoustic intensity using the cross-spectral density of two microphone signals," *The Journal of the Acoustical Society of America*, vol. 62, no. 4, pp. 1057–1059, Oct. 1977. [Online]. Available: https://doi.org/10.1121/1.381601
[59] H. Huang *et al.*, "A method of speech periodicity enhancement using transform-domain signal decomposition." *Speech Commun*, vol. 67, pp. 102–112, Mar 2015.
[60] J. C. Segura *et al.*, "Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust asr," in *INTERSPEECH*, 2002.
[61] G. Parikh *et al.*, "The influence of noise on vowel and consonant cues," *The Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3874–3888, Dec. 2005. [Online]. Available: https://doi.org/10.1121/1.2118407
[62] S. L. Salzberg, "C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann Publishers, inc., 1993," *Machine Learning*, vol. 16, no. 3, pp. 235–240, Sep. 1994. [Online]. Available: https://doi.org/10.1007/bf00993309
[63] "TIMIT Acoustic-Phonetic Continuous Speech Corpus," (Accessed on: 21-11-2022). [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1
[64] "NOIZEUS: A noisy speech corpus for evaluation of speech enhancement algorithms," (Accessed on: 21-11-2022). [Online]. Available: https://ecs.utdallas.edu/loizou/speech/noizeus/
[65] D. Pearce, "Aurora speech recognition experimental framework," (Accessed on: 21-11-2022). [Online]. Available: http://aurora.hsnr.de/index-2.html
[66] P. C. M. Wong *et al.*, "Cortical mechanisms of speech perception in noise," *J. Speech Lang. Hear. Res.*, vol. 51, no. 4, pp. 1026–1041, Aug. 2008.
[67] Vesper, "VM1010," (Accessed on: 21-11-2022). [Online]. Available: https://vespermems.com/products/vm1010/