Voice Activity Detection System for Smart Earphones

Narimene Lezzoum, Ghyslain Gagnon and Jérémie Voix

Abstract — *This paper presents a real-time voice activity detection (VAD) algorithm implemented in a miniature Digital Signal Processor (DSP) for in-ear listening devices such as earphones or headphones. This system allows consumers to hear external speech signals such as public announcements or oral communication while listening to music without removing their listening devices. The proposed algorithm uses two normalized energy features that compare the energy in the frequency region containing speech information with the frequency regions typically containing noise. The extraction of the normalized features represents the key of the proposed VAD since it eliminates the need for a signal-to-noise ratio (SNR) estimator. The VAD's decision is made using two threshold comparison rules computed from the normalized features and a hangover scheme triggered after a given number of observations. The algorithm parameters, namely the frequency regions' boundaries, number of observations, two decision thresholds and hangover's duration, have been optimized off-line using a genetic algorithm. The performance of the proposed VAD is compared to a benchmark algorithm in four noise environments and three SNRs. Results show that the average false positive rate (FPR) of the proposed algorithm is 4.2% and the average true positive rate (TPR) is 91.4 % compared to the benchmark algorithm which has a FPR average of 29.9 % and a TPR average of 79.0 %. The proposed VAD is implemented in hardware to validate its reliability and complexity*[1].

Index Terms — Smart earphones, voice activity detection, energy based feature, real-time algorithm, digital signal processor.

I. Introduction

Nowadays, smart-phones, mp3 players, and other portable audio player devices are ubiquitous.

Wearing earphones or headphones for listening to music in public places such as airports, airplanes, or railway stations causes sensory and cognitive distractions and isolates the wearer from the external environment. For example, in a railway station when train departures are announced, earphone wearers may miss this announcement and consequently miss their train. Similarly, in an airplane, passengers must remove their earphones when a steward is addressing them.

To palliate problems caused by the wearing of earphones in public places, several tools have been developed to enable consumers to hear external signals, ranging from push-to-hear electronic devices to dedicated wireless systems.

Earphone manufacturers have developed systems which include a microphone and a push button that allows the users to mute the music and transmit external sounds to the ear, thus allowing communication without the need to remove the earphones. These devices are either available as external dangles or included directly into the headphones. Since the users must manually push a button, they must know that a spoken message is addressed, which is unsuitable in situations where no visual cue is available (public announcement, for example).

Software tools for external signals transmission are also available in smart-phones. They enable consumers to hear the external environment while listening to the music when the loudness of the external environment exceeds a certain threshold. Although these tools let the earphone wearers remain aware of their external environment, they can be annoying since all signals (useful and not-useful) are transmitted to the ear whenever they reach the predetermined loudness threshold.

Sophisticated wireless systems have also been developed to address this problem. These systems transmit the announcements to the wearer's audio device via a network, and then play relevant announcements in the earphones [1]. This method requires a specific infrastructure in a given location, and the user cannot benefit from this technology where the infrastructure has not been developed.

The present paper describes a real-time Voice Activity Detection (VAD) system for smart earphones that can be integrated to current advanced communication earpieces [2]. The proposed system discriminates between a speech (useful) signal and noise (not-useful) signal to transmit speech signals through the earphones while blocking noise signals. A miniature Digital Signal Processor (DSP) is integrated in the earphones for real-time speech and noise discrimination.

Voice activity detection is commonly used in various speech-based applications. In voice over IP transmission and GSM communication, a VAD is used to encode non-speech segments with a lower bit rate than speech segments and thus reduce the transmission rate [3]. It is also widely used in

[1] This work was supported by the *Sonomax-ETS Industrial Research Chair in In-Ear Technologies* (CRITIAS).

N. Lezzoum is with the Department of Mechanical Engineering, Ecole de technologie superieure, University of Quebec, H3C 1K3, Canada (e-mail: narimene.lezzoum@ens.etsmtl.ca).

G. Gagnon is with the Department of Electrical Engineering, Ecole de technologie superieure, University of Quebec, H3C 1K3, Canada (e-mail: ghyslain.gagnon@etsmtl.ca).

J. Voix is with the Department of Mechanical Engineering, Ecole de technologie superieure, University of Quebec, H3C 1K3, Canada (e-mail: jeremie.voix@etsmtl.ca).

Contributed Paper
Manuscript received 10/01/14
Current version published 01/09/15
Electronic version published 01/09/15.

human/machine interaction applications [4], [5] for speech recognition or speaker identification and verification to reduce false alarm rates due to the use of noise segments in the recognition process. Likewise, VADs are used for noise reduction in hearing aids [6] and recently for smart hearing protection [7].

The performance of VADs relying solely on the extraction of one or several features [3], [8], [9] degrades when the signal-to-noise ratios (SNR) decreases [10]. To palliate this problem, other VADs have been developed and rely on the estimation of the *a posteriori* and *a priori* SNR using the signal first frames, assumed to contain only noise signals [11]. Unfortunately, these VADs become sensitive to changes in the SNR [12]. Learning techniques or modeling algorithms have also been applied to VADs [13], [14] making the VAD efficient but more complex and difficult to implement in a DSP with limited hardware resources for real-time applications.

Recently, Hsu *et al* [15] proposed an energy-based VAD where the decision is made using a threshold upon the energy of the frequency modulation of harmonics. This VAD has shown its effectiveness in low SNRs and requires low computational resources. However its response delay makes it unsuitable for real-time low-latency applications.

While a relatively low-complexity VAD has been proposed based on the inter-quartile range statistic feature [7], the current approach proposes improvements, using simpler energy-based features, for an efficient implementation in a low-power DSP.

The proposed VAD is implemented in a miniature DSP for smart earphones or headphones applications. The proposed solution can be integrated into active noise control headphones, which are already equipped with external microphone and other electronics. It can even be retrofitted to traditional headphones or earphones by integrating a miniature external microphone and DSP.

This paper is organized as follows: Section II presents the proposed smart earphones and their operating principle. Section III describes the proposed VAD algorithm. In Section IV, the parameters used in the VAD's decision are defined and their off-line optimization using a genetic algorithm is performed. Section V presents the validation of the proposed VAD, and Section VI describes its implementation in a low-power DSP and its real-time validation in the embedded system. Finally, Section VII concludes the paper.

II. The Smart Earphone

Smart earphones are traditional earphones, in which a field-programmable electronic hardware is embedded (Fig.1). To capture signals, a miniature external microphone is connected to the audio input of an ultra-low power DSP. The DSP output is connected to a miniature loudspeaker to transmit the desired signals to the ear [16].

The main task of the proposed system is the discrimination between speech and noise signals to allow speech signals to get through the earphones while blocking noise signals when speech is absent to enable the wearer to listen to music. Fig.2 illustrates the operating principle of the whole system.

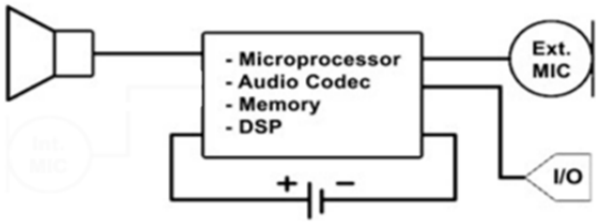

Fig. 1. The hardware resources embedded in the smart earphones.

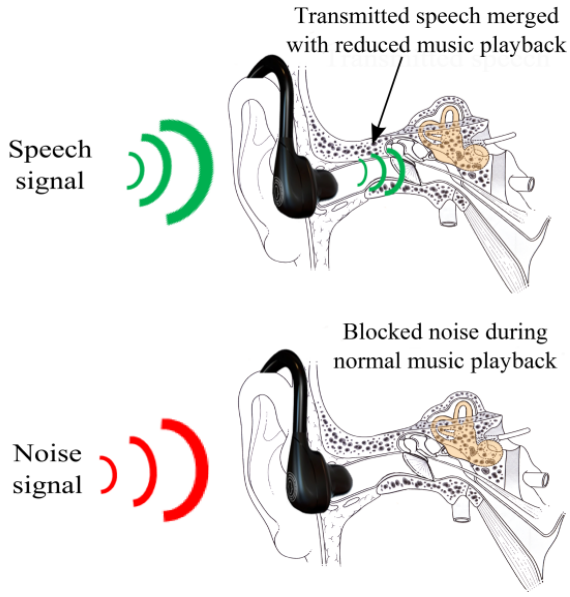

Fig. 2. The selective operating principle of the system.

III. The Proposed VAD Algorithm

A study conducted by Parikh *et al* [17] on the influence of noise on vowels and consonants concluded that when the speech signal is corrupted by noise, the first formant can be reliably detected compared to the second formant, which is heavily masked by noise in low SNR (0 dB). Based on these findings, we propose the use of an energy feature which is extracted from the frequency region containing the first formant for speech characterization. Thereafter, this feature is normalized using two noise features extracted from the frequency regions containing typical noise information. The normalization of the energy feature eliminates the need for an SNR estimator.

The VAD's decision is made after multiple observations using two decision thresholds, determined from the normalized energy features in addition to a hangover scheme to consider the "long time" information, knowing that the speech signal is highly time-correlated [18]. The value of the two thresholds, the frequency bounds, the number of

observations and the hangover parameters are optimized off-line using a genetic algorithm. The optimization increases the performance of the proposed VAD by maximizing the F1 score [19].

Fig.3 illustrates the detailed architecture of the proposed VAD algorithm. The signal is first time-windowed into *i* frames. Features are extracted and the decision *D(i)* is made after *N* observations based on two thresholds and a hangover scheme.

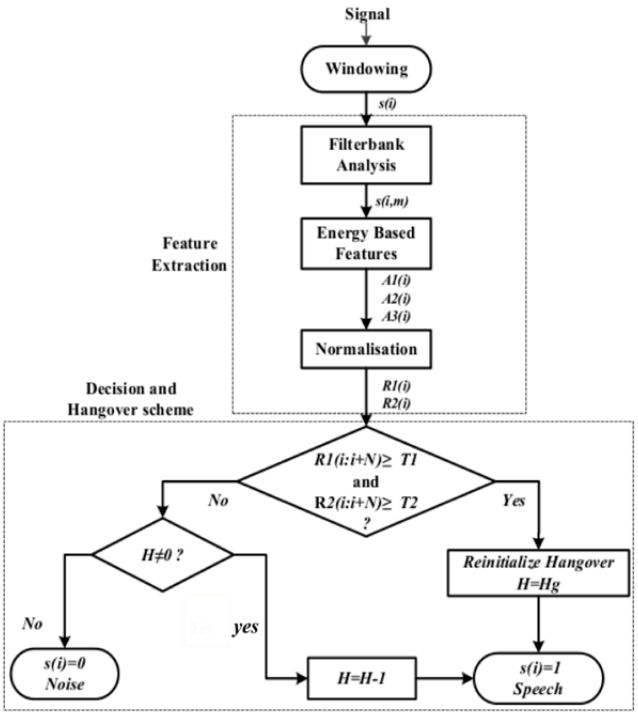

Fig. 3 Block diagram of the proposed VAD algorithm.

A. Windowing

The entire signal is cut into frames using a Hamming window. The length of each frame is 25 ms with 80 % overlap.

B. Feature Extraction

1) Filter Bank

The incoming signal is filtered into *M=3* frequency bands using 4th order Butterworth filters. Cut-off frequencies of the 3 bands (15-153 Hz, 153-1323 Hz, 1323-1944 Hz) have been optimized off-line using a Genetic algorithm (see section IV).

2) Energy-based Features

Parikh *et al* [17] concluded that when the speech signal is corrupted by noise; the first formant can be reliably detected. Based on the conclusions of this study, the energy of each frequency band is calculated.

Fig.4 illustrates an example of the energy in the three frequency bands for one speech frame produced by a male speaker corrupted by car noise with 10, 5 and 0 dB SNRs. *A1, A2, A3* denote the energy in the first, second, and third frequency bands respectively. One can see that in the second frequency band, which contains the first formant of the speech frame (a voiced phoneme), the energy of the speech is significantly higher than the energy of the noise in this band, whereas the energy of the noise in the first band (especially in 0 dB SNR) is higher for noise than speech.

3) Normalization

While Fig.5 shows that *A2* is a reliable indicator of the presence of speech, it cannot be used directly with a decision threshold in the VAD because it is dependent on the input signal level. Thus, the following normalized ratios, which increase the VAD's performance by taking advantage of the different frequency content of speech and noise (*A1* and *A3*), are proposed:

$$R1 = \frac{A2}{A1} \quad (1)$$

$$R2 = \frac{A2}{A3} \quad (2)$$

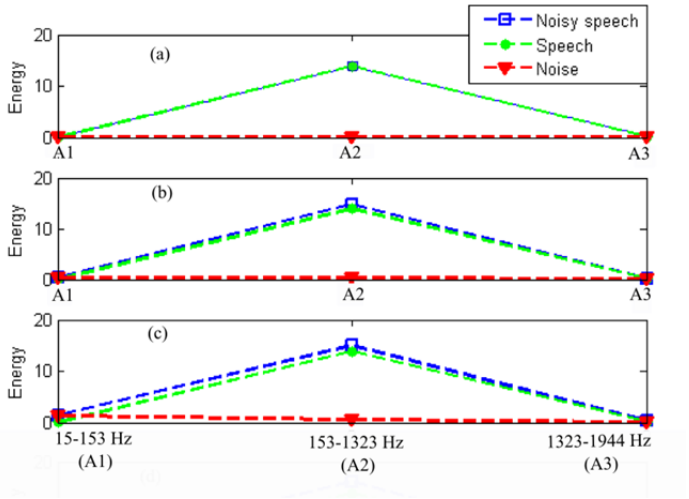

Fig. 4 Energy in three frequency bands for one signal frame with (a) 10 dB, (b) 5 dB, and (c) 0 dB SNR.

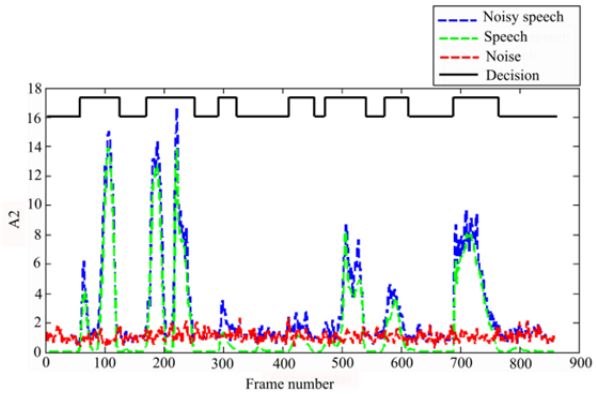

Fig. 5 A2 in speech, noise, and a noisy speech signal with 0 dB SNR, in addition to the hand-labeled decision on clean speech.

R1 is normalized by the low-frequency components, knowing that noise signals generally have more energy in the lower frequencies than speech signals [20]. *R2* is normalized

by the high-frequency components that characterize high frequency noise signals.

The VAD's decision is based on ratios *R1* and *R2,* thus eliminating the need for an SNR estimator.

C. VAD's Decision

1) The decision thresholds

Two decision thresholds *T1* and *T2* are fixed upon the ratios *R1* and *R2*.

The VAD's decision is made after *N* observations:

$$D(i) = \begin{cases} 1 & \text{if } \mathrm{R1}(i:i+N) \geq \mathrm{T1} \text{ and } \mathrm{R2}(i:i+N) \geq \mathrm{T2} \\ 0 & \text{else} \end{cases} \quad (3)$$

with *N* being the number of consecutive observations, *i* the frame number and *D(i)* the decision in the current frame.

2) Start and End of Speech Confirmation Parameters

The VAD's decision is made after multiple observations (start of speech confirmation parameter). These observations are defined by the number of consecutive frames having ratios *R1* and *R2* higher than thresholds *T1* and *T2* respectively and after which the decision is to be set to 1 (speech). Ramirez et *al.* [21] demonstrated that taking several frames into account in the VAD improves the reliability of its decision.

In the proposed VAD, the number of consecutive frames should not exceed 8 frames to not exceed a delay of 40 ms.

Hangover schemes (end of speech confirmation parameter) have been widely used in VADs to reduce the false rejection rate attributable to the non-detection of low energy speech frames containing consonants such as fricatives and unvoiced stops [11] [18].

In the adaptive Multi-Rate (AMR) VAD [9], the hangover was set to 2 seconds if the signal is of a complex nature.

IV. Off-Line Parameter Optimization

The choice of the two decision thresholds *T1* and *T2* depends on the desired specificity and sensitivity of the VAD. High decision thresholds make the VAD more specific than sensitive, which minimizes both the False Positive Rate (FPR) and True Positive Rate (TPR). Low decision thresholds make the VAD more sensitive by maximizing the TPR and FPR.

The two decision thresholds *T1* and *T2*, the number of consecutive frames (start of speech confirmation), and the hangover (end of speech confirmation), in addition to the frequency bands' boundaries are optimized off-line using a genetic algorithm approach by maximizing an objective function.

A. Objective Function

In the literature, VAD performance evaluation can be performed using various metrics [10]. Nevertheless, solving an optimization problem requires the use of one metric reflecting the entire performance of the VAD algorithm. For this purpose, the F1 score is used as the objective function [18]:

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \quad (4)$$

with:

$$\mathrm{precision} = \frac{\mathrm{TPR}}{\mathrm{TPR} + \mathrm{FPR}} \quad (5)$$

$$\mathrm{recall} = \frac{\mathrm{TPR}}{\mathrm{TPR} + \mathrm{FNR}} \quad (6)$$

The F1 score combines the TPR, FPR and False Negative Rate (FNR). It reflects the VAD's accuracy by considering its precision and recall.

In VAD algorithms, TPR, FPR and FNR are respectively: the ratio of speech frames classified as speech, the ratio of noise frames classified as speech, and the ratio of speech frames classified as noise. In the existing VAD algorithms, these rates are calculated in noisy speech signals to distinguish between speech and noise frames. However, for a smart earphone application, the TPR and FNR are calculated for noisy speech signals and the FPR for noise signals. This evaluation method focuses on the fact that once the speech frames have been detected, the detection of the next non-speech frames does not have any detrimental effect on the performance of the proposed VAD. Whereas when no speech signal is present, the detection of noise frames and their transmission to the protected ear is significantly detrimental on the performance of the proposed VAD.

B. Audio Signals used for the Off-line Optimization

Off-line parameters optimization is conducted to maximize the F1 score, using a small number of noisy speech signals. In the envisioned application, noise signals typical of everyday environments are to be used. Thus 20 speech signals (14 speech signals produced by male speakers and 6 speech signals produced by female speakers) from the TIMIT database [22] corrupted by "Airport" noise recorded in real world environment with 5 dB SNR are used. Speech and noise were artificially mixed together with 5 dB SNR.

The TIMIT database was chosen for the envisioned application because the speech signals in this database are not altered by filters such as the ITU MIRS or ITU G.712, that tend to consider the realistic frequency characteristics of terminals and equipment in the telecommunication area [23].

C. Genetic Algorithm for Off-Line Parameters Optimization

Genetic algorithms [24] are randomized search and optimization techniques based on the mechanism of natural selection and natural genetics. They are robust and efficient, they adapt to a wide variety of environments and they produce a near optimal solution when solving an optimization problem.

The genetic algorithms are used to optimize the frequency bands' boundaries, the hangover, the number of consecutive observations, and the decision thresholds.

In the optimization process, the lower and upper frequency bounds variations for the three band-pass filters are illustrated in Table I. The lower bound of the second and third frequency bands correspond to the upper bound of the first and second frequency bands respectively.

TABLE I
FREQUENCY BANDS' LOWER AND UPPER BOUNDS FOR THE OPTIMIZATION PROCESS

Bounds	Lower bound (Hz)	Upper bound (Hz)
1	10	20
2	50	250
3	250	1500
4	1500	6000

The hangover varies from 50 to 300 frames with a step of one frame (0.25 to 1.5 second). The number of observations varies from 4 to 8 consecutive frames, which is equivalent to a decision delay varying from 20 to 40 ms.

After 10 generations, the genetic algorithm reached an optimal solution with an F1 score of 98.5 %. Fig.6 shows a plot of the function's best and mean penalty values in each generation with each generation being composed of 40 individuals. The optimization process gave a hangover value of Hg=1.26 seconds and a number of consecutive frames N=7. The optimized cut-off frequencies for the three band-pass filters are: [15, 153, 1323, 1944] Hz.

These parameters are then used for the decision-making and the validation of the proposed VAD algorithm using a validation database.

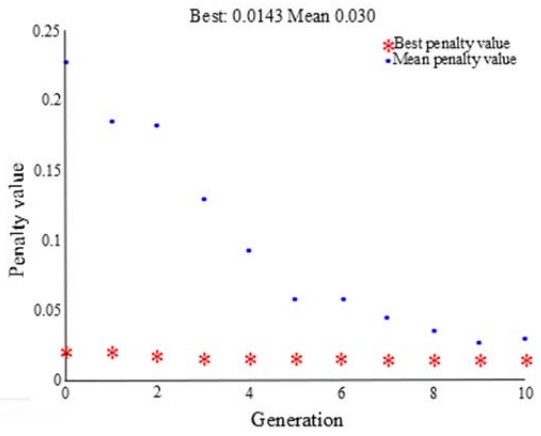

Fig. 6 Penalty values (1-F1) of the optimization process using Genetic Algorithm.

V. EXPERIMENTS AND VALIDATION

A. Validation database

The validation database is composed of 10 sentences produced by 630 speakers (439 male speakers and 191 female speakers) from the TIMIT database [22]. Signals are sampled at 8 kHz. All 10 sentences are concatenated into one signal.

Noisy speech signals were created by adding the same noise at three SNRs (10, 5, and 0 dB) to each concatenated speech signal.

Four noise signals obtained from real world recordings were used. These noises are representative of everyday environments to which consumers may be exposed to:

- *Car*: this environment tends to mimic the noise of the wind perceived by car passengers with opened windows.
- *Airport*: this noise was recorded in the hall of an airport, with talking crowds and baggage trolleys passing by.
- *Hammer*: this noise contains transient noises. It is used to mimic some scenarios such as renovations in the neighborhood, or constructions in the street.
- *Train:* this noise was recorded near a railway with sounds of trains passing by.

B. Performance Evaluation

The performance evaluation is conducted using the F1 score, in addition to the TPR and FPR. The proposed algorithm is compared to Sohn's VAD [11] which uses the first signal's frames to estimate the *a posteriori* and the *a priori* SNR to make the decision.

Fig.7 illustrates the F1 score results of both algorithms in all noise conditions. As it can be seen in this figure, the F1 score of the proposed algorithm outperforms the F1 score of Sohn's algorithm in all noise environments and SNRs.

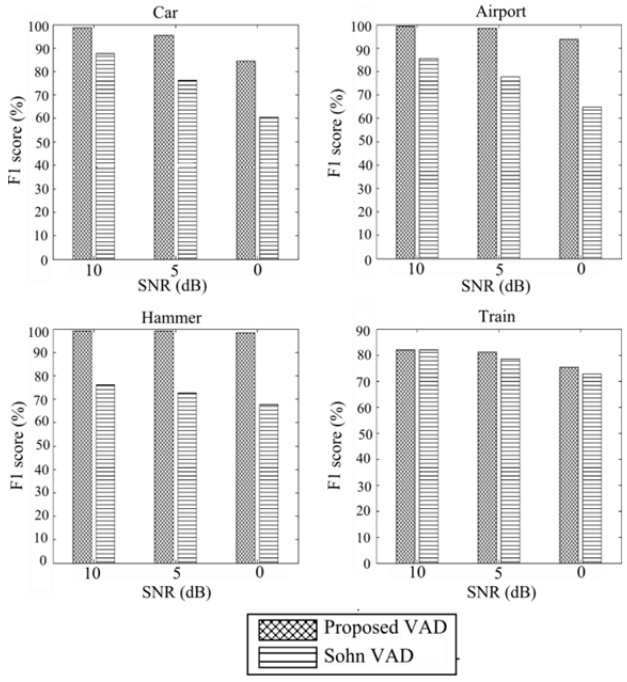

Fig. 7 F1 scores of Sohn's and the proposed VAD in four noise environments with 10, 5, and 0 dB SNR.

TABLE II
PERFORMANCE EVALUATION OF THE PROPOSED AND SOHN'S VADS IN FOUR NOISE ENVIRONMENTS AND THREE SNRS.

Noise Environment		Proposed VAD (%)		Sohn VAD (%)	
Noise	SNR	TPR	FPR	TPR	FPR
Car	10 dB	97.6	0	87.5	20.9
	5 dB	91.3	0	76.1	20.9
	0 dB	73.4	0	60.2	20.9
Airport	10 dB	98.4	0	85.5	14.9
	5 dB	97.0	0	72.9	14.9
	0 dB	88.4	0	55.6	14.9
Hammer	10 dB	98.7	0	91.7	50.2
	5 dB	98.4	0	85.3	50.2
	0 dB	96.7	0	77.6	50.2
Train	10 dB	97.7	16.8	92.7	33.9
	5 dB	91.1	16.8	86.4	33.9
	0 dB	68.4	16.8	76.5	33.9
Average	Average	91.4	4.2	79.0	29.9

In applications such as the smart earphones (to simultaneously enable the wearer to listen to music and transmit speech signals when present), the less desirable situation is the detection of short-time noise. This situation occurs when the false positive rate is high. Table II presents the true positive rate and false positive rate for the two VADs.

The FPR average of the proposed VAD is 4.2 % compared to Sohn's VAD which has a FPR average of 29.9 %. The same FPR is found in the three SNRs of each noise since both VADs are insensitive to the level of the incoming signal.

Furthermore, the TPR of the proposed algorithm is higher than the TPR of Sohn's algorithm in all noise environments in the range of 5 and 10 dB SNR. This is due to the hangover scheme presented previously, which permits the detection of almost all the speech frames without interruptions or mid-speech clipping.

VI. HARDWARE IMPLEMENTATION

A. DSP Overview

The DSP used for the implementation of the VAD is a stream-oriented DSP core provided in a small 32-lead, 5 mm x 5 mm package. The Analog to Digital Converter (ADC) and the Digital to Analog Converter (DAC) are high quality 24 bit stereo audio converters, and can operate at sampling frequencies ranging from 8 kHz to 96 kHz. The DSP core consist of a simple multiply-accumulate (MAC) unit with a data source and a coefficient source. Three RAMs are encompassed in the address space of the DSP: the program RAM, the coefficient RAM, and the data RAM.

The program RAM governs the execution of the instructions in the core, and cannot exceed 1024 instructions per audio frame. The parameter RAM stores the initial coefficients of the program and cannot exceed 1024 coefficients, while the data RAM stores audio data-words for processing in addition to some run-time parameters. The data RAM is divided into two memory addressing types: modulo and non-modulo memories. Each of the modulo and non-modulo data RAM offer 4096 memory words.

B. Hardware VAD Implementation

The Auditory Research Platform (ARP) [25] integrates the DSP in addition to other associated electronics such as audio inputs, audio outputs, and battery. It is used to implement the proposed VAD in real-time. Fig.8 illustrates the ARP platform with two earpieces, in each earpiece an external miniature microphone and an internal miniature loudspeaker are integrated for external sound acquisition and VAD's decision transmission respectively.

Fig. 8 The auditory research platform in which the VAD is implemented for real-time processing connected to two earpieces for audio signal acquisition and VAD's decision transmission.

The hardware VAD implementation is made following the steps described in Section III.

The resulting number of instructions per audio frame is 890, which is equivalent to a rate of 87 % from the entire program RAM. The data RAM used by the VAD is 346 (8 % from the entire modulo data RAM, and 0 % from the non-modulo data RAM), while the coefficient RAM used is 240 (23 % of the coefficient RAM).

C. VAD real-time tests and Validation

The real-time validation of the proposed algorithm is performed using some of the noisy audio samples used in the first validation process presented in section V.

For this purpose, the audio input of the ARP was connected to the audio output of a computer in which the noisy signals were playing, while the output of the VAD's decision was saved in the computer to compare it with the result of the first validation presented in section V.

Fig.9 illustrates an example of the comparison between the results of the VAD before its hardware implementation and the VAD implemented in the DSP. This comparison is made using a signal composed of 3 ms of airport noise, 3 ms of speech corrupted by airport signal with 5 dB SNR, and 3 ms of airport noise signal. The decision of the VAD in simulation and its hardware implementation are equivalent.

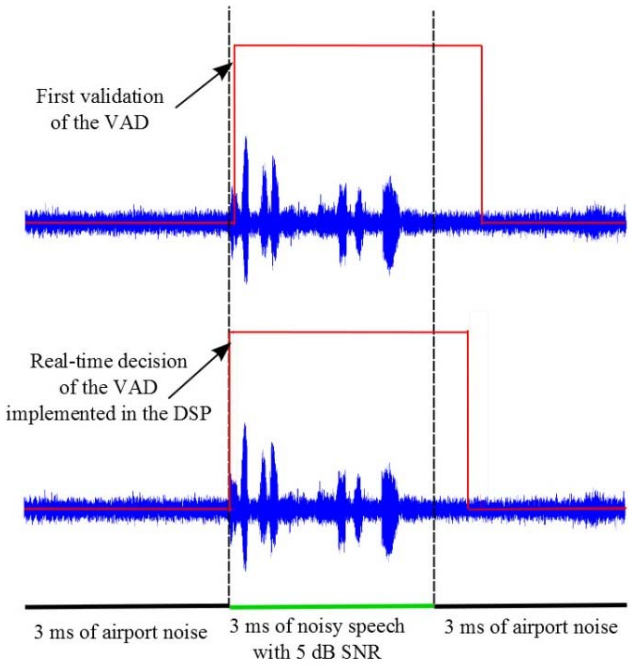

Fig. 9 Comparison between the VAD decision on the computer and the real-time VAD decision obtained from the output of the DSP.

VII. Discussions and Conclusions

In this paper, a robust and yet simple real time VAD for smart earphones is presented. This VAD uses an energy-based feature for the characterization of speech and noise signals. The speech and noise characteristics are thereafter normalized and two decision thresholds are determined. The decision is made after multiple observations and triggers a hangover scheme.

The algorithm parameters are optimized off-line using a genetic algorithm by maximizing the F1 score which represents the global performance of the VAD. The parameters optimization is performed using 20 speech signals corrupted by airport noise with 5 dB SNR.

The first experiment for the validation of the proposed VAD was conducted using 10 sentences produced by 439 male speakers and 191 female speakers corrupted by four noise environments. These experiments showed that the proposed VAD is more efficient than a benchmark VAD. Coupling multiple observations and the hangover scheme in the decision process shows that the proposed VAD detects almost all speech signals without interruption since the true positive rate average is 91.4 %. The entire VAD system was validated for the smart earphones application. The proposed VAD was implemented in a miniature low-power DSP integrated in a research platform in which the audio inputs, battery, and other electronics were selected for real-time implementation. The hardware resources show that other tasks can be combined to the VAD such as a low complexity on-line parameters optimization algorithm to allow to the VAD to adapt for each noise environment in which the smart earphones are used.

References

[1] V. R. Desai “Smart tool for headphones”, US patent, 2014.
[2] J. Voix, N. Laperle, J. Mazur, A. Bernier, “Advanced communication earpiece device and method”, US patent, 2014
[3] ITU, T, "Annex B: Silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70", International Telecommunication Union, 1996.
[4] N. Cho and E.k.Kin, "Enhanced voice activity detection using acoustic event detection and classification, " *IEEE Trans. Consumer. Electron.,* vol. 57, no. 1,pp. 196-202, 2011.
[5] H. Lee, S. Chang, D. Yook, and Y. Kim, "A voice trigger system using keyword and speaker recognition for mobile devices", IEEE *Trans. Consumer. Electron*., vol. 55, no.4, pp.2377-2384, 2009.
[6] K. Chung, "Challenges and recent developments in hearing aids: part I. speech understanding in noise, microphone technologies and noise reduction algorithms, " *Trends in Amplification*, vol. 8, no. 3, pp. 83-124, 2004.
[7] N. Lezzoum, G. Gagnon, J. Voix, "A low complexity voice activity detector for smart hearing protection of hysperacusic persons", in Proc. Interspeech, Lyon, France, pp 723-727, 2013.
[8] R. Tucker, "Voice activity detection using a periodicity measure," in IEE Proc. Communications, Speech and Vision, 1992.
[9] ETSI, "Digital cellular telecommunications system (Phase 2+); Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) speech traffic channels; General description (GSM 06.94 version 7.1.0 Release 1998)," 1999.
[10] F. Beritelli, S. Casale, G. Ruggeri, and S. Serrano, "Performance evaluation and comparison of G.729/AMR/fuzzy voice activity detectors," *IEEE Signal Process. Lett*., vol. 9, no. 3, pp. 85-88, 2002.
[11] J. Sohn, N. S. Kim, and W. Sung, "A Statistical model-based voice activity detection," *IEEE Signal Process. Lett*., vol. 6, no. 1, pp. 1998-2000, 1999.
[12] M. H. Moattar and M. M. Homayounpour, "A simple but efficient real-time voice activity detection algorithm," Proc. European Signal Processing Conference, Glasgow, Scotland, pp 2549-2553, 2009.
[13] J. Wu and X. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Process. Lett.*, vol. 18, no. 8, pp. 466-469, 2011.
[14] X. Liu, Y. Liang, Y. Lou, H. Li, and B. Shan, "Noise-robust voice activity detector based on hidden semi-markov models," in Proc. International Conference on Pattern Recognition, Istanbul, Turkey, pp 81-84, 2010.
[15] C.-C. Hsu, T.-E. Lin, J.-H. Chen, and T.-S. Chi, “Voice activity detection based on frequency modulation of harmonics,” in Proc. *IEEE* International Conference on Acoustics Speech and Signal Process. 2013, Vancouver, Canada, pp. 6679–6683, 2013.
[16] M.-A. Carbonneau, N. Lezzoum, J. Voix, and G. Gagnon, "Detection of alarms and warning signals on an digital in-ear device," *International Journal of Industrial Ergonomics*, pp. 1-9, 2012.
[17] G. Parikh and P. Loizou, "The influence of noise on vowel and consonant cues," *The Journal of the Acoustical Society of America*, vol. 118, no. 6, pp. 3874-3888, 2005.
[18] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio, Speech, Language Process*., vol. 14, no. 2, pp. 412-424, 2006.
[19] V. Rijsbergen, *Information retrievial*, Butterworths, 1979.
[20] H. Levitt, "Noise reduction in hearing aids: a review," *Journal of Rehabilitation Research and Development*, vol. 38, no. 1, pp. 111-121, 2001.
[21] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using multiple observation likelihood ratio Test," *IEEE Signal Process. Lett*., vol. 12, no. 10, pp. 689-692, 2005.
[22] S. V. Zue and J. Glass, "Speech Database Development: TIMIT and Beyond," *Speech Communication*, vol. 9, no. 4, pp. 351-356, 1990.
[23] P. D. Hirsch Hans-günter, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Automatic Speech Recognition : Challenges for the Next Millennium", 2000.
[24] D. E. Goldberg, *Genetic Algorithms in search, optimization, and machine learning*, Addison-Wesley, 1989.
[25] K. Mazur, J. Voix, “Implementing 24-hour in-ear dosimetry with recovery” in Proc. International Conference on Acoustics, New York, USA, 2013.

BIOGRAPHIES

Narimene Lezzoum received the Master degree in electrical engineering from University of Science and Technology Houari Boumedienne, Algiers, Algeria, in 2010. She is currently a PhD candidate within the Industrial Research Chair in In-Ear Technologies at École de technologie supérieure (ÉTS), Montreal, QC, Canada. Her current research interests include speech and signal processing for real world embedded applications.

Ghyslain Gagnon received the B.Eng. and M.Eng. degrees in electrical engineering from École de technologie supérieure, Montreal, Canada in 2002 and 2003 respectively. He also received the Ph.D. degree in electrical engineering from Carleton University, Canada in 2008. From 2003 to 2004, he worked for ISR Technologies where he designed and implemented several critical synchronization modules for a software-defined radio which later obtained the editors' choice award in 2007 by the portable design magazine. He is now an associate professor with the department of Electrical Engineering, École de technologie supérieure. He is inclined towards industrial research partnerships. His research aims at mixed-signal circuits and systems, as well as digital signal processing.

Jérémie Voix received a B.Sc in Fundamental Physics from Lille 1 University (France) in 1995, a M.A.Sc in Acoustics from Université de Sherbrook in 1997 and a PhD -with great distinction- from École de technologie supérieure (ÉTS) in Montreal in 2006. From 2001 to 2010, he is VP of Scientific Research than CTO at Sonomax Hearing Healthcare. Since 2010, he is Associate Professor at ÉTS and leads the Sonomax-ÉTS Industrial Research Chair in In-Ear Technologies. Together with a motivated and gifted team, he is working to merge hearing aid, hearing protection and communication technologies within a single in-ear device. Research horizons include interpersonal radio-communication systems and on-board hearing health monitoring.