LETTER

Efficient Implementation of Statistical Model-Based Voice Activity Detection Using Taylor Series Approximation

Chungsoo LIM[†], **Soojeong LEE**[††], **Jae-Hun CHOI**[††], ***Nonmembers***, ***and*** **Joon-Hyuk CHANG**[††a)], ***Member***

SUMMARY In this letter, we propose a simple but effective technique that improves statistical model-based voice activity detection (VAD) by both reducing computational complexity and increasing detection accuracy. The improvements are made by applying Taylor series approximations to the exponential and logarithmic functions in the VAD algorithm based on an in-depth analysis of the algorithm. Experiments performed on a smartphone as well as on a desktop computer with various background noises confirm the effectiveness of the proposed technique.

key words: *voice activity detection, Taylor series approximation, embedded systems*

1. Introduction

Voice activity detection (VAD) has a wide variety of applications such as speech coding, speech recognition, noisy speech enhancement, hands-free conference, and echo cancelation [1]. For this reason, VAD has been extensively studied, and various types of VAD algorithms have been proposed to improve the voice activity detection accuracy [1]–[3]. While detection accuracy is still of paramount importance, the computational complexity of VAD algorithms also deserves attention because these algorithms are usually executed on a short frame basis on battery-powered embedded systems in which power consumption is critically important. Even for alternating current (AC)-powered systems, reducing energy consumption is also desirable. Therefore, computational complexity reduction of VAD algorithms is investigated to increase their power efficiency without degrading detection accuracy.

There are many different VAD algorithms; in this study, we target the statistical model-based VAD with predicted *a priori* signal-to-noise ratio (SNR) estimation because it is well-known for its superior detection accuracy [1]. Among many subroutines in the VAD algorithm, we observe that the exponential and logarithmic functions are executed as many times as the product of the number of considered frequency bins and the number of input frames; these functions are responsible for about 15% of the VAD algorithm's total execution time.

Because this represents an ample opportunity for improvement in overall execution time, and because the prevailing modular programming style allows these functions to be modified easily, we decide to simplify these functions by replacing them with their Taylor series representations [4]. However, because detection decisions are directly dependent on these functions, a naive application of the Taylor approximation may degrade the detection accuracy considerably. Hence, herein we describe a judicious application of the Taylor series approximation that not only maintains but actually improves the detection accuracy of the statistical model-based VAD, based on careful design and detailed analysis of this VAD algorithm.

Manuscript received September 9, 2013.
Manuscript revised November 15, 2013.

[†]The author is with Korea National University of Transportation, Chungju-si, Rep. of Korea.
[††]The authors are with Hanyang University, Seoul Rep. of Korea.
a) E-mail: jchang@hanyang.ac.kr

DOI: 10.1587/transfun.E97.A.865

2. Target VAD Algorithm

In this section, the targeted VAD algorithm is briefly reviewed and analyzed. First, we assume that a noise signal $d(t)$ is added to a speech signal $x(t)$, with their sum being denoted by $y(t)$, that is

$$y(t) = x(t) + d(t) \tag{1}$$

Taking the discrete Fourier transform (DFT) gives us

$$Y_k(n) = X_k(n) + D_k(n) \tag{2}$$

where k is the frequency bin index ($k = 0, 1, \ldots, L-1$) and n is the frame index ($n = 0, 1, \ldots$). Assuming that speech is degraded by uncorrelated additive noise, the two hypothesis H_0 and H_1, which indicate speech absence and presence in the noisy spectral component $Y_k(n)$, respectively, are given by

$$H_0 : \text{speech absent} : Y_k(n) = D_k(n) \tag{3}$$

$$H_1 : \text{speech present} : Y_k(n) = X_k(n) + D_k(n) \tag{4}$$

With the complex Gaussian probability density functions (pdf's) assumption [2], the distributions of the noisy spectral components conditioned on each of the hypotheses are given by

$$p(Y_k(n) \mid H_0) = \frac{1}{\pi\lambda_{d,k}(n)} \exp\left\{-\frac{|Y_k(n)|^2}{\lambda_{d,k}(n)}\right\} \tag{5}$$

$$p(Y_k(n) \mid H_1) = \frac{1}{\pi(\lambda_{d,k}(n) + \lambda_{x,k}(n))} \exp\left\{-\frac{|Y_k(n)|^2}{\lambda_{d,k}(n) + \lambda_{x,k}(n)}\right\} \tag{6}$$

where $\lambda_{x,k}(n)$ and $\lambda_{d,k}(n)$ denote the variance of $X_k(n)$ and $D_k(n)$, respectively. The likelihood ratio of the kth frequency bin is derived as

$$\Lambda_k(n) \equiv \frac{p(Y_k(n)|H_1)}{p(Y_k(n)|H_0)} = \frac{1}{1+\xi_k(n)} \exp\left\{\frac{\gamma_k(n)\xi_k(n)}{1+\xi_k(n)}\right\} \quad (7)$$

where $\xi_k(n) \equiv \lambda_{x,k}(n)/\lambda_{d,k}(n)$ and $\gamma_k(n) \equiv Y_k(n)/\lambda_{d,k}(n)$ are called the *a priori* and *a posteriori* SNRs, respectively [2]. The *a posteriori* SNR $\gamma_k(n)$ is obtained by $\lambda_{d,k}(n)$, which is updated during periods of speech absence, and the *a priori* SNR $\xi_k(n)$ is estimated based on the predicted (PD) estimation as follows [5]:

$$\hat{\xi}_k(n) \equiv \frac{\hat{\lambda}_{x,k}(n)}{\hat{\lambda}_{d,k}(n)} \quad (8)$$

where $\hat{\lambda}_{x,k}(n)$ and $\hat{\lambda}_{d,k}(n)$ are the estimates for $\lambda_{x,k}(n)$ and $\lambda_{d,k}(n)$, respectively. These estimates are computed as

$$\begin{aligned}\hat{\lambda}_{d,k}(n+1) &= \zeta_d \hat{\lambda}_{d,k}(n) + (1-\zeta_d) E\left[|D_k(n)|^2 | Y_k(n)\right] \\ \hat{\lambda}_{x,k}(n+1) &= \zeta_x \hat{\lambda}_{x,k}(n) + (1-\zeta_x) E\left[|X_k(n)|^2 | Y_k(n)\right] \end{aligned} \quad (9)$$

where ζ_x(=0.98) and ζ_d(=0.99) are the smoothing parameters under a general stationary assumption on $D_k(n)$ and $X_k(n)$. The expectations in the above equations can be described as

$$\begin{aligned} &E\left[|D_k(n)|^2 | Y_k(n)\right] \\ &\quad = E\left[|D_k(n)|^2 | Y_k(n), H_0\right] P(H_0|Y_k(n)) \\ &\qquad + E\left[|D_k(n)|^2 | Y_k(n), H_1\right] P(H_1|Y_k(n)) \end{aligned} \quad (10)$$

$$\begin{aligned} &E\left[|X_k(n)|^2 | Y_k(n)\right] \\ &\quad = E\left[|X_k(n)|^2 | Y_k(n), H_0\right] P(H_0|Y_k(n)) \\ &\qquad + E\left[|X_k(n)|^2 | Y_k(n), H_1\right] P(H_1|Y_k(n)) \end{aligned} \quad (11)$$

where the speech absence probability is given by [5]

$$p(H_0|Y_k(n)) = \frac{1}{1 + \frac{P(H_1)}{P(H_0)} \Lambda_k(n)} \quad (12)$$

where $P(H_0)(= 1 - p(H_1))$ is the *a priori* probability of speech absence.

From (8) to (10), it can be seen that estimating the *a priori* SNR $\xi_k(n+1)$ requires the LR $\Lambda_k(n)$. This relationship will be used in the next section to explain the strategy for applying Taylor series approximation. Finally, the decision rule for VAD is formulated as the geometric mean of the LRs computed for the individual frequency bins such that

$$\log \Lambda(n) = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda_k(n) \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad (13)$$

with L being the total number of frequency bins and η denoting the threshold for detection. As can be seen from (7) and (13), the VAD algorithm includes one exponential function and one logarithmic function, which are executed for every frequency bin in each input frame, and the VAD decisions are directly computed with the help of these functions.

3. Strategy for Applying Taylor Series Approximation to VAD

3.1 Target Function Rearrangement

Whereas approximating the exponential function in (7) with a Taylor polynomial is straightforward, approximating the logarithmic function in (13) requires a well-thought-out strategy. In order to approximate the logarithmic function in (13) with a Taylor polynomial, we firstly devise a better way of applying the approximation than applying it directly to the logarithmic function. Instead of using $\Lambda_k(n)$ from (7) to compute $\log \Lambda_k(n)$, we first take the logarithm of (7) and then use the resulting equation, which is given by

$$\log \Lambda_k(n) = \log\left(\frac{1}{1+\xi_k(n)}\right) + \frac{\gamma_k(n)\xi_k(n)}{1+\xi_k(n)} \quad (14)$$

If $\Lambda_k(n)$ were used only for the VAD decision, using (14) to compute $\log \Lambda_k(n)$ would be natural and more efficient because the exponential function in (7) can be removed. However, because $\Lambda_k(n)$ is also used to estimate the *a priori* SNR as explained in the previous section, the use of (7) is still required. Therefore, using $\Lambda_k(n)$ to obtain $\log \Lambda_k(n)$ can be considered a natural step, but we nonetheless adopt (14) to calculate $\log \Lambda_k(n)$ for the following two reasons. First, the input to the logarithmic function $(1/(1+\xi_k(n)))$ is limited to the interval between zero and one, allowing us more freedom in choosing a Taylor polynomial that satisfies our purpose, which will be described shortly. Second, the use of (14) prevents the errors that are introduced by the approximated version of the exponential function in (7) from significantly affecting the VAD decisions. That is, if $\Lambda_k(n)$ is used to compute $\log \Lambda_k(n)$, the VAD accuracy could be severely degraded due to the serial use of the two approximations.

3.2 Target Series Selection

Basically, the Taylor series expansion for the exponential function is represented as follows [4]:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \quad \forall x \quad (15)$$

Note that, on the other hand, there exist a few Taylor series for the natural logarithmic function according to the input value range. In our case, where the range is confined between zero and one, there are three candidates (see Fig. 1):

$$\text{A.} \quad \log x = \sum_{n=1}^{\infty} \frac{1}{n} \left(\frac{x-1}{x}\right)^n \qquad x \geq \frac{1}{2} \quad (16)$$

$$\text{B.} \quad \log x = \sum_{n=1}^{\infty} \frac{(1-(-1)^n)}{n} \left(\frac{x-1}{x+1}\right)^n \qquad x > 0 \quad (17)$$

$$\text{C.} \quad \log x = \sum_{n=1}^{\infty} \left(-\frac{1}{n}\right)(1-x)^n \qquad 0 < x \leq 2 \quad (18)$$

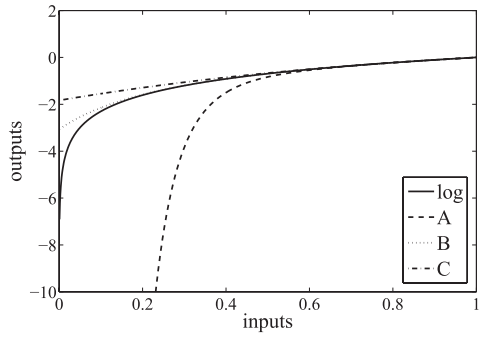

Fig. 1 Comparison of the three Taylor series approximations (all of Taylor series order 3) to the original log function.

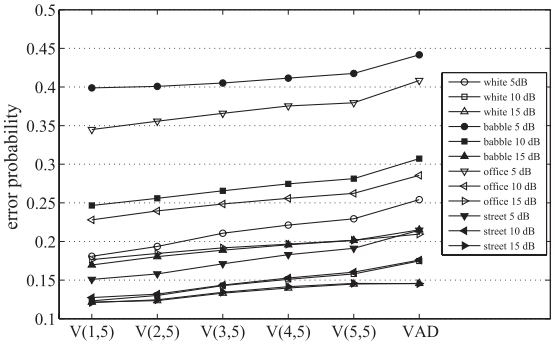

Fig. 2 VAD error probability for different Taylor series orders in the logarithmic approximation function. V(x, y) refers to VAD using the logarithmic approximation of order x and the exponential approximation of order y; VAD refers to the original algorithm without approximation.

Among these candidates, we set our own criteria for choosing the one that would be used to approximate the log function within the framework of the statistical model-based VAD. The first criterion is computational complexity. Since the main purpose of using approximations is to reduce the computational complexity of the original function, the chosen approximation should be as simple as possible. As far as this criterion is concerned, (18) is the most suitable because the others include an additional division, which is the most expensive single arithmetic operation.

The second criterion is approximation characteristic. Although high approximation accuracy is generally preferred, we have an unconventional requirement: the approximation should be larger than the original log function for small inputs (i.e., inputs between 0 and 0.3), but it should be accurate for larger inputs. This requirement is based on the observation that it is more probable for speech frames to generate more inputs (to the log function) between 0 and 0.3 than nonspeech frames. This is simply because speech frames inherently have higher *a priori* SNR than nonspeech frames. For example, according to our experimental results, 23% of the inputs to the log function from speech frames fall into this range, whereas only 10% of the inputs from nonspeech frames do. Due to this disparity in the input distribution to the log function between speech frames and nonspeech frames, having an approximation that is larger than the log function in this particular range, such as B and C in Fig. 1, increases $\log \Lambda_k(n)$ of speech frames more frequently than that of nonspeech frames. And, because increased $\log \Lambda_k(n)$ implies higher $\log \Lambda(n)$ in (13), the number of beneficial switches from incorrect VAD decisions as nonspeech to correct VAD decision as speech is more than the number of detrimental conversions from correct VAD decision as nonspeech to incorrect VAD decisions as speech, thus improving VAD accuracy overall. Although both B and C satisfy this criterion, we decide to use C because C is computationally less expensive and larger than B for small inputs.

4. Experimental Results

The effectiveness of the proposed technique was evaluated with 456 s of speech data recorded by four males and four females. We manually labeled each 10 ms frame to provide reference decisions [1]; the proportions of voiced, unvoiced, and silent frames were 44.5%, 13.4%, and 42.1%, respectively. To create various noise environments, white, babble, office, and street noises were added to the clean speech data with SNRs of 5, 10, and 15 dB.

To measure the execution time of the VAD, we prepared a Linux desktop computer equipped with a 3.0 GHz quad-core x86 processor as well as a smartphone powered by a 1.5 GHz dual-core ARM processor. To measure the influence of the approximation degree on VAD accuracy, we varied the Taylor-series orders for both exponential and logarithmic functions from one to five. Figure 2 shows how the order of the approximation for the logarithmic function affects the VAD accuracy, which was measured as VAD error probability (the sum of the false alarm and missing probabilities) [1]. We fixed the order of the exponential approximation to five to isolate the influence of the approximated logarithmic function. As the order increased, the VAD error probabilities also elevated irrespective of noise types and SNRs. Recalling the benefits of overestimating $\log \Lambda_k(n)$ for inputs between 0 and 0.3, the observed behavior is attributable to the fact that this overestimation diminishes as the Taylor-series order increases and the approximation function thereby resembles the original logarithmic function more closely. However, it should be noted that even the fifth-order Tayor series, the highest tested, outperformed the original VAD in every case tested except when street noise was added with 15 dB SNR.

Similarly, we examined the impact of the Taylor-series order for the exponential function on the VAD accuracy, holding constant the first-order logarithmic approximation that performed best (Fig. 3). In contrast to the logarithmic case, the exponential function resulted in fewer errors as it was approximated more accurately. For example, using the first-order and fifth-order Taylor series respectively for the logarithmic and exponential functions reduced the error probability by 20.93% on average relative to the use of the original, unapproximated functions.

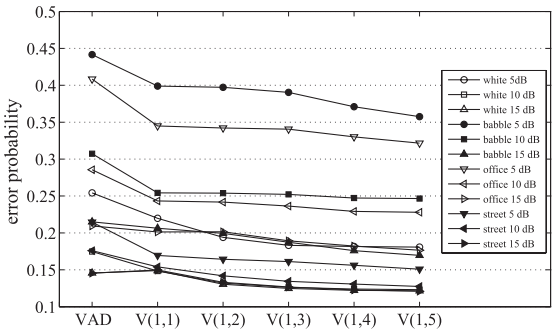

Fig. 3 VAD error probability for different Taylor series orders in the exponential approximation function. VAD refers to the original algorithm without approximation; V(x, y) refers to VAD using the logarithmic approximation of order x and the exponential approximation of order y.

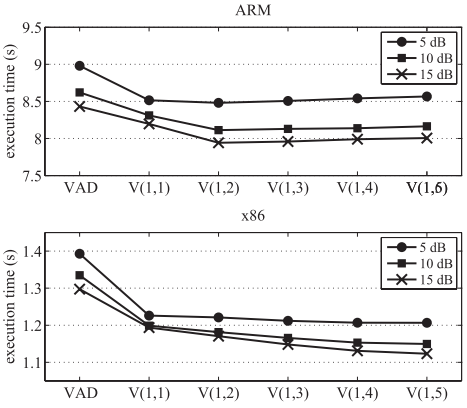

Fig. 4 VAD execution time comparison of different Taylor series orders on the x86 and ARM processors.

To measure the reductions in execution time enabled by the use of the simpler Taylor series alternatives, we conducted experiments with various Taylor series orders on two different platforms: a Linux desktop and an Android smartphone. The resulting execution times are shown in Fig. 4. Note that the results in the figure are the average values calculated over all noise types. The Taylor series order for the logarithmic function was fixed at one because the VAD execution time was rather insensitive to the Taylor series order for the logarithmic function due to ample instruction level parallelism in the VAD algorithm that can hide a few additional arithmetic operations required for a higher order Taylor series approximation of the logarithmic function.

As can be seen from the figure, replacing the exponential and the logarithmic functions with their respective Taylor polynomials decreases the VAD latency for both processors, but the two processors produce slightly different behaviors: the execution times decrease as the Taylor series order for the exponential function increases on the x86 processor, but on the ARM processor, the execution times are minimized at V(1,2) and increase with Taylor series order thereafter. This small disparity was observed because the execution time decreases due to higher SNRs estimated by higher Taylor series orders (see the figure for this behavior) outweighed the execution time increase caused by additional instructions of higher orders approximations only on the x86 processor, which could execute the additional instructions more efficiently.

The maximum execution time reduction was measured to be 13.55% on the x86 processor and 5.75% on the ARM processor. For the x86 processor, the greatest execution time reduction and the highest VAD accuracy concurred for the same orders, V(1,5). However, for the ARM processor, the most accurate algorithm was also V(1,5) but the fastest was V(1,2). Nonetheless, the use of the most accurate algorithm, V(1,5), allowed an similar execution time reduction of 4.99%.

5. Conclusions

We have proposed a Taylor series-based technique that effectively enhances both VAD accuracy and latency. In contrast to the general belief that approximations lead to accuracy degradations, the proposed technique can achieve both goals through judicious application of Taylor-series approximation to the VAD algorithm. The proposed technique was proven effective in extensive experiments including various noises types and intensities on two representative real-world hardware platforms, showing its suitability to potentially be adopted in future VAD implementations.

Acknowledgement

This research was supported by the MSIP, Korea, under the ITRC support program supervised by the NIPA (NIPA-2013-H0301-13-4005) and this work was supported by NRF grant funded by the MEST (NRF-2011-0009182).

References

[1] J.-H. Chang, N.S. Kim, and S.K. Mitra, "Voice activity detection based on multiple statistical models," IEEE Trans. Signal Process., vol.56, no.6, pp.1965–1976, June 2006.
[2] J. Sohn, N.S. Kim, and W. Sung, "A statistical model-based voice activity detection," IEEE Signal Process. Lett., vol.6, no.1, pp.1–3, Jan. 1999.
[3] S.-K. Kim and J.-H. Chang, "Voice activity detection based on conditional MAP criterion incorporating the spectral gradient," Signal Process., vol.92, no.7, pp.1699–1705, July 2012.
[4] E.T. Whittaker and G.N. Watson, A course of modern analysis, Cambridge University Press, 2009.
[5] J.-H. Chang and N.S. Kim, "Speech enhancement: New approaches to soft decision," IEICE Trans. Inf. & Syst., vol.E84-D, no.9, pp.1231–1240, Sept. 2001.