A DUAL-MICROPHONE SUBBAND-BASED VOICE ACTIVITY DETECTOR USING HIGHER-ORDER CUMULANTS

Elias Nemer and Ashutosh Pandey

Broadcom Corporation
5300 California Ave
Irvine, CA, 92617
{enemer, pandeya}@broadcom.com

ABSTRACT

The paper proposes a robust dual-microphone algorithm for Voice Activity Detection (VAD) suitable for detecting speech arriving from random directions. The algorithm is based on the use of higher order statistics (HOS) in the complex subband domain in order to effectively detect voicing segments and distinguish them from nonharmonic noise. Metrics based on new established properties of the 2nd and 4th-order cumulants of complex exponentials are derived. The pros and cons of each of these are analyzed and validated through simulation in various SNR conditions. The results show the proposed scheme is effective in discriminating voiced speech segments, and is robust to Gaussian-like and real-life recorded noises, even in low SNR.

Index Terms— VAD, HOS, cumulants, multi-mics

1. INTRODUCTION

Voice Activity Detection (VAD) refers to the process of classifying an audio recording into speech and non-speech segments. This ability to distinguish active speech from noise is an important feature in a number of audio applications, such as speech recognition, speech enhancement, speech coding, and echo cancellation. Over the past several years, a number of VAD approaches have been presented in the literature, some of which are based on statistical models [1], cepstrum coefficients [2], entropy [3], or various other time frequency metrics.

The ongoing challenge for VAD algorithms is to make them resilliant to noisy and interfering environments which are prevalent in typical audio applications. To improve the detection accuracy in low SNR, various schemes such as pattern recognition [4], adaptive energy thresholds [5], third order statistics [6], and periodicity estimators [7] were proposed.

In general, single-channel VADs are practical for close-talking applications. However, in distant-talking contexts, they can become unreliable due to environmental noises, speech attenuation, and reverberation. In today's applications, such as interactive TVs and hands-free mobile phones, multichannel VADs are needed because they exploit the spatial information provided by multiple sensors. The VADs based on spatial correlation or homogeneity of the DOA [8][9] are good at detecting directional sound sources from various locations, though they cannot discriminate between speech and various noise sources from different DOA.

In [10], we proposed a robust single-channel VAD algorithm based on established properties of the 3rd and 4th- order statistics, when considering the flat spectrum of the LPC residual. In this paper, we extend these ideas to a multichannel case, and the complex subband context. We establish new properties for the HOS of complex exponentials and use these to discriminate voicing energy in the subband structure. Metrics are developed at the subband and full-band levels to yield a robust way to distinguish –mostly voiced- speech from non-harmonic noise segments.

The paper is structured as follows: Section 2 derives analytical expressions for the 4th-order cumulants. Section 3 discusses the metrics for voicing detection. Section 4 illustrates the overall algorithm. Section 5 provides the simulation results and section 6 the conclusion.

2. HOS OF COMPLEX SINUSOIDS

In the following, some key properties relating the 2nd and 4th-order cumulants of complex sinusoidal signals –with complex amplitude and random phase- are discussed.

2.1 The 4th-Order Cumulant as Function of the 2nd- Order

Theorem: The Kurtosis, or 4th -order cumulant at lag 0 of a complex harmonic signal, is nonzero and can be expressed as a function of the 2nd order statistics of the signal.

Proof:

From the general expression of the 4th-order cumulant:

$$C^4_{z_1 z_2 z_3 z_4} = E[z_1 z_2 z_3 z_4] - E[z_1 z_2]E[z_3 z_4] - E[z_1 z_3]E[z_2 z_4] - E[z_1 z_4]E[z_2 z_3] \quad (1)$$

where z_1, z_2, z_3, z_4 represent time samples of the same signal (separated by a given lag) or samples from different signals. We set the identities as:

$$x_1 \equiv z_1 = z_3 \qquad x_1^* \equiv z_2 = z_4 \quad (2)$$

and obtain the expression of the 4th-order cumulant (at lag zero):

$$\begin{aligned} C^4_{x_1} = {} & E[x_1(n)x_1^*(n)x_1(n)x_1^*(n)] \\ & - E[x_1(n)x_1^*(n)]E[x_1(n)x_1^*(n)] \\ & - E[x_1^2(n)]E[x_1^{*2}(n)] \\ & - E[x_1(n)x_1^*(n)]E[x_1^*(n)x_1(n)] \end{aligned} \quad (3)$$

Or simply:

$$C^4_{x_1} = E[|x_1(n)|^2|x_1(n)|^2] - 2(E[|x_1(n)|^2])^2 - E[x_1^2(n)]E[x_1^{*2}(n)] \quad (4)$$

For the case of a complex harmonic signal with random phase of the form:

$$x_1 = a_1 e^{-j(\omega_1 n + \varphi)}$$

It is straightforward to show that $E[|x_1(n)|^2] = |a_1|^2$,

$E[|x_1(n)|^2|x_1(n)|^2] = |a_1|^4$, and $E[x_1^2(n)] = [x_1^{*2}(n)] = 0$
Thus, the 4th-order cumulant becomes -using (4):

$$C_{x_1}^4 = -|a_1|^4 \qquad (5)$$

yielding the relation between the 2nd and the 4th-orders as:

$$C_{x_1}^4 = -\{E[|x_1(n)|^2]\}^2 = -\{C_{x_1}^2\}^2 \qquad (6)$$

Therefore, the 4th-order cumulant at lag 0 (or kurtosis) of a harmonic signal can be written as a function of the squared energy (or 2nd order cumulant) of the signal. The above derivation can be extended to the case of two or more harmonics and yield similar results.

2.2 Cross-Cumulant of Delayed Exponential Signals

Theorem: The cross-cumulant between two complex harmonic signals separated by a time delay L_0 (Figure 1) reaches a maximum negative value when the correlation lag L.matches the time delay.

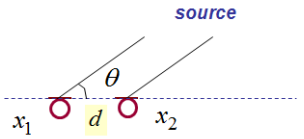


Figure 1 : Signal model for 2-mic input.

Proof:
The signal from the two microphones can be written –in the transform domain- as a scaled and delayed version of the source:

$$X_1(n) = \alpha S_n = Ae^{j(wn+\varphi)}$$
$$X_2(n) = \beta S_{n-L_0} = Be^{j(w(n-L_0)+\varphi)}$$

The cross-cumulant between the two signals at a lag L is:

$$C_{X_1X_2}^4(L) = E[X_1^2(n)X_2^{*2}(n+L)] - E[X_1^2(n)]E^*[X_2^2(n+L)] - 2(E[X_1(n)X_2^*(n+L)])^2 \qquad (7)$$

Thus, given the following identities, for the case of the complex exponential signals:

$X_1^2(n) = A^2e^{j2wn}$, $X_2^2(n) = B^2e^{j2w(n-L_0)}$,
$X_2^{*2}(n) = B^{*2}e^{-j2w(n-L_0)}$, $X_2^*(n) = B^*e^{-jw(n-L_0)}$, and
$X_2(n+L) = Be^{jw(n-L_0+L)}$.

Substituting into (7), the first term in the cross-cumulant is:

$$E[X_1^2(n)X_2^{*2}(n+L)] = A^2B^{*2}e^{-j2w(L-L_0)} \qquad (8)$$

The second term is:

$$E[X_1^2(n)]E^*[X_2^2(n+L)] = 0 \qquad (9)$$

The third term is:

$$E[X_1(n)X_2^*(n+L)] = A\cdot B^*\cdot e^{-jw(L-L_0)} \qquad (10)$$

Combining the terms into (7) yields the expression:

$$C_{X_1X_2}^4(L) = -A^2B^{*2}e^{-j2w(L-L_0)} \qquad (11)$$

Thus, the cross cumulant reaches a maximum (negative) value when the lag matches the time delay: $L = L_0$.

Corollary 1: The magnitude of the normalized cross-cumulant between two delayed complex harmonic signals is one.

Proof:
To eliminate the effect of signal energy, a normalized cross-cumulant can be deduced by normalizing it with the individual channels' cumulants as:

$$Norm\_C_{X_1X_2}^4(L) = \frac{C_{X_1X_2}^4(L)}{\sqrt{C_{X_1}^4(0)C_{X_2}^4(0)}} \qquad (12)$$

Using equations (11) and (5), the magnitude of the ratio (12) becomes:

$$\left|Norm\_C_{X_1X_2}^4(L)\right| = \frac{\left|C_{X_1X_2}^4(L)\right|}{\sqrt{C_{X_1}^4(0)C_{X_2}^4(0)}} = \frac{\left|-A^2B^{*2}e^{j2w(L_0-L)}\right|}{\sqrt{(-|A|^4)(-|B|^4)}} = \left|-e^{j2w(L_0-L)}\right| = 1$$

Corollary 2:
The general relation between the cross-correlation and cross-cumulant of two complex signals for any lag L is given by:

$$C_{X_1X_2}^2(L) \equiv E[X_1(n)X_2^*(n+L)] = \sqrt{-C_{X_1X_2}^4(L)} \qquad (13)$$

Proof:
The relation is straightforward from equations (10) and (11).

3. VOICE DETECTION BASED ON HOS

The speech from each microphone is sampled at 16 kHz, and the signal is divided into complex subbands using a polyphase filter bank. A total of 48 bands are used, thus each subband may contain 0, 1, or 2 harmonics, depending on pitch.

3.1 Subband-Based Voicing Metrics

We introduce two ratios for voicing detection: the first involves 4th-order cumulants only. The second involves a combination of 2nd and 4th-order cumulants. The following are the conditions for voicing in each subband:

a. Relation between the 2nd and 4th-order cumulants: The kurtosis of each channel is negative and its absolute value is greater than the energy (2nd moment) of that channel:

$$C_{x_1}^4 < -C_{x_1}^2 \qquad C_{x_2}^4 < -C_{x_2}^2 \qquad (14)$$

b. Relation between the individual 4th-order cumulants and the cross-cumulant of the two channels. The following ratio is near unity (for any lag L or frequency band k):

$$\left|Norm\_C_{X_1X_2}^4(L)\right|_{\text{band-k}} = \left|\frac{\left|C_{X_1X_2}^4(L)\right|}{\sqrt{C_{X_1}^4(0)C_{X_2}^4(0)}}\right|_k \approx 1 \qquad (15)$$

Theoretically, this metric is not susceptible to noise since both the numerator and denominator involve only higher cumulants. However, since the two are zero for Gaussian type noise, the ratio may take unpredictable values during non-speech. Practically, the variance of the estimators for the numerator and denominator prevents the zero division, though the value will fluctuate with noise level.

c. Relation between the 4th-order cross-cumulant and the 2nd order cross-correlation between the two channels. The following ratio is near unity (for any lag L or frequency band k):

$$\left|C_{X_1X_2}^4\_C_{X_1X_2}^2(L)\right|_{\text{band-k}} = \left|\frac{\sqrt{-C_{X_1X_2}^4(L)}}{C_{X_1X_2}^2(L)}\right|_k \approx 1 \qquad (16)$$

Theoretically, this metric is affected by the noise due to the 2nd order term in the denominator, thus we would expect the ratio to deviate from 1 in low SNR and weak speech segments. Practically, the variance of the estimator of the 4th-order cumulant (in the numerator) is also a function of the noise energy [11],

therefore during noise, the ratio value will depend on the noise level. To mitigate this effect, both ratios (15, 16) are computed whenever the condition given by (14) is met. A plot of the two ratios is given in Figure 2 for an SNR of 5 dB for a given frequency subband.

We deduce that the two ratios are robust to the presence of noise. While both ratios exhibit the overall unity value to some degree, the ratio of the normalized cross-cumulants (Eq 15) is more accurate and has fewer fluctuations than that of the 2nd and 4th-order cumulants (Eq 16) in all SNR conditions.

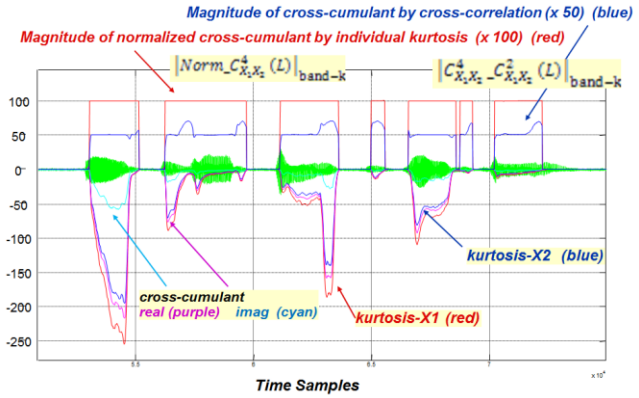


Figure 2 : VAD metrics, subband #3, SNR = 5 dB.

3.2 Frame-Based Voicing Metrics

In order to yield a frame-based (i.e., a full band) voicing detector, similar metrics to (15) and (16) are derived by summing the numerators and denominators in the two ratios across the frequency subbands:

a. Full-band normalized cross-cumulant:

$$\left|Norm\_C^4_{X_1X_2}(L)\right|_{\text{Frame}} = \frac{\Sigma_k \left|C^4_{X_1X_2}(L)\right|_k}{\Sigma_k \sqrt{C^4_{X_1}(0) C^4_{X_2}(0)}\Big|_k} \qquad (17)$$

b. Fullband cross-cumulant normalized by cross-correlation:

$$\left|C^4_{X_1X_2} - C^2_{X_1X_2}(L)\right|_{\text{Frame}} = \frac{\Sigma_k \left|\sqrt{-C^4_{X_1X_2}(L)}\right|_k}{\Sigma_k \left|C^2_{X_1X_2}(L)\right|_k} \qquad (18)$$

The above metrics are expected to be near unity when computed as shown since the subbands that contain harmonics will contribute to both the numerator and denominator, and those that have no harmonic energy will not add anything to either.

c. A final metric is considered and consists of summing the ratio of equation (15) across all subbands:

$$\sum_k \left|Norm\_C^4_{X_1X_2}(L)\right|_k = \sum_k \frac{\left|C^4_{X_1X_2}(L)\right|}{\sqrt{C^4_{X_1}(0) C^4_{X_2}(0)}}\Bigg|_k \qquad (19)$$

A plot of all three frame-based metrics for the case of 5 dB SNR is shown in Figure 3 below.

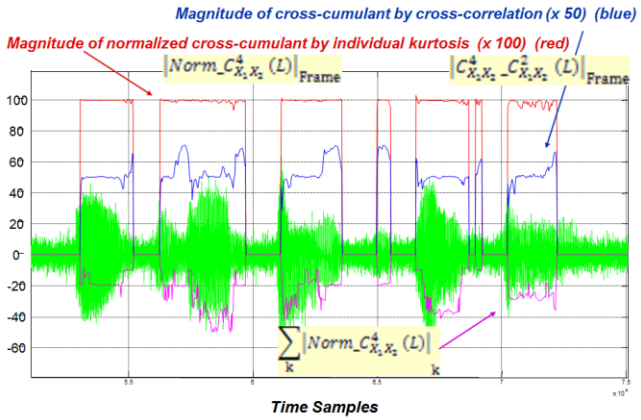


Figure 3 : Frame-based VAD metrics; SNR = 5 dB.

As in the case of subband-based metrics, we find the normalized cross-cumulant (Eq 17) to be accurate and very robust to noise. It does, however, miss the onset of speech syllables and trailing edges. The first is due to the averaging history delay while the second is due to the trailing edge of certain phonemes being non-harmonic. The same holds for unvoiced segments, which go mostly undetected.

4. EXPERIMENTAL RESULTS

4.1 HOS-based VAD Algorithm

The overview of the algorithm is given in Figure 4 below. The signals from both channels are divided into complex subbands using polyphase filter banks. The 2nd and 4th-order statistics are computed for each channel and for the cross-channels. The ratios (Eq 15, 16, 17, 18, and 19) are computed for the bands and averaged over the frame. Thresholds are used and the decision is made to declare the bands and/or the frame as containing voicing energy or not.

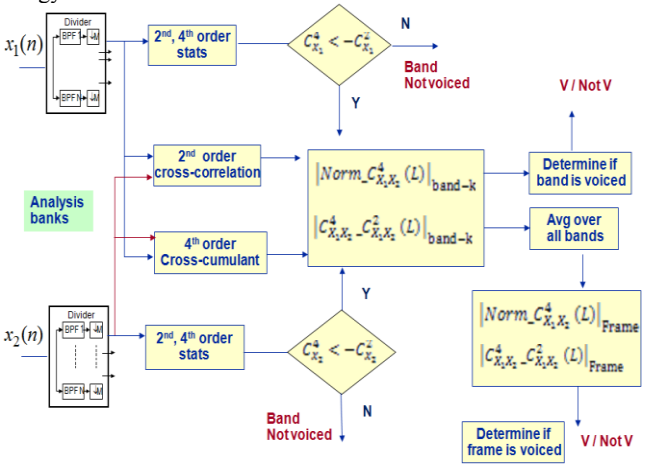
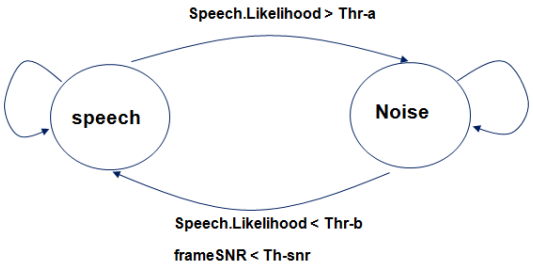


Figure 4 : VAD algorithm overview.

A frame likelihood measure is computed, based on the normalized fullband cross-cumulant (Eq 17):

$$\text{Speech. Likelihood}_{\text{Fr}} = \alpha e^{-\beta\left|1-\left|C^4_{X_1X_2} - C^2_{X_1X_2}(0)\right|_{\text{Fr}}\right|} \qquad (20)$$

Where α and β are tuning parameters. An HOS-based VAD state machine is constructed as follow:



4.2 Comparative Analysis with multi-mic VAD Algorithms

4.2.1 A Dual-Mircophone Gaussian-based VAD

The statistical detector developed by Sohn [12][13] is based on modeling the spectral components of speech and noise as complex Gaussian distributions, from which a likelihood ratio is deduced. The approach hinges on a proper estimate of the a posteriori and a priori SNR defined defined by Ephraim *et al*.[14]. To extend the algorithm to a multi-microphone, the noise suppression approach and its corresponding multi-channel SNR's proposed in [15] is used.

4.2.2 A DOA-based VAD

The algorithm proposed in [9] uses the homogeneity of the DOA values estimated from the phases across the frequency bins of the FFT transforms of both microphone signals. From the estimates, a measure of the entropy is computed, from which the VAD decision are deduced.

4.2 Results on collected noisy speech data

Clean speech is recorded in an un-echoic chamber using two microphones spaced apart by 10 cm, with male and female speakers, and phonetically balanced sentences. Noise is either generated synthetically (Gaussian) or recorded separately from a street corner or in a cafeteria using the same microphone pairs. Noisy speech with various SNR's is generated. The detection results of the full-band VAD metric are compared with the hand-labeled speech/non-speech as well as the voiced speech traces. The probability of correctly detecting speech, voiced speech, as well as that of false classification are listed in Tables 1 to 3 below.

Table 1 : HOS-based VAD simulation results.

	SNR	Pc Speech	Pc Voiced segment	Prob. False classification
Gaussian	10 dB	0.753	0.8719	0.114
	5 dB	0.6394	0.7985	0.12
Bable	15 dB	0.762	0.8836	0.1294
	5 dB	0.5802	0.7649	0.2216
	0 dB	0.6364	0.8148	0.2894
Wind noise	10 dB	0.7848	0.8875	0.14

Table 2 : Statistical VAD simulation results.

	SNR	Pc Speech	Pc Voiced segment	Prob. False classification
Gaussian	10 dB	0.776	0.9458	0.0692
	5 dB	0.5045	0.6152	0.262
Bable	15 dB	0.997	0.9999	0.368
	5 dB	0.9119	0.9798	0.3735
	0 dB	0.6608	0.7188	0.3277
Wind noise	10 dB	0.9807	0.9998	0.454

Table 3 : DOA-based VAD simulation results.

	SNR	Pc Speech	Pc Voiced segment	Prob. False classification
Gaussian	10 dB	0.7204	0.89	0.1554
	5 dB	0.6715	0.8082	0.129
Bable	15 dB	0.973	0.9915	0.3175
	5 dB	0.8759	0.9692	0.3891
	0 dB	0.7404	0.8341	0.35
Wind noise	10 dB	0.6039	0.6699	0.381

Both the statistical and DOA-based VADs quickly skew towards speech frames in moderate level babble noise, resulting in high false detection. The HOS-based VAD has a more graceful degradation in the various noise types. The DOA-based VAD has a poor performance in wind noise, even though it performs very well in Gaussian noise. The performance of the statistical VAD degrades for low SNR in Gaussian noise, and in general, it proves to be very sensitive to threshold settings.

5. CONCLUSION

We proposed a two-microphone VAD algorithm based on the 2nd and 4th-order cumulants of the complex subband domain signals. The derived metrics are based on newly established properties that higher order statistics of complex exponentials would exhibit and the fact that they are different from those of Gaussian noise. Simulation results from synthetic and recorded noise confirm the effectiveness of the metrics used. Comparison to recently developed multi-channel VAD algorithms show the proposed method is effective in a variety of real-like recorded noises.

The algorithm is particularly robust in detecting voiced segments in various noises, even in very low SNR. Because of this feature, it is deemed valuable in applications where detection of most, though not all, of the voiced segments is needed, such as in direction of arrival or in long-term pitch estimation. Finally, the proposed scheme can be extended to any number of microphone pairs.

6. REFERENCES

[1] O. Gauci, C. J. Debono and P. Micallef. “A Maximum Log-Likelihood Approach to Voice Activity Detection”, in *Proc. ISCCSP 2008*. pp:383-387.

[2] J.A. Haigh and J.S. Mason, “Robust Voice Activity Detection Using Cepstral Features”, in *Proc. of IEEE TENCON'93*, vol.3, pp. 321-324, 1993, Beijng.

[3] Kun-Ching Wang and Yi-Hsing Tasi. “Voice activity detection algorithm with low signal-to-noise ratios based on spectrum entropy”. 2008 *Second International Symposium on Universal Communication*. pp:423-428.

[4] F. Beritelli, S. Casale, and A. Cavallaro, “A robust voice activity detector for wireless communications using soft computing,” *IEEE J. Sel. Areas Commun*., vol. 16, no. 9, pp. 1818–1829, Dec. 1998.

[5] H. Ozer and S. G. Tanyer, “A geometric algorithm for voice activity detection in nonstationary Gaussian noise,” in *Proc. EUSIPCO*, Rhodes, Greece, Sep. 1998.

[6] Hui-jing Dou Zhao-yang Wu Yan Feng Yan-zhou Qian, “Voice Activity Detection Based on the Bispectrum,” in *Proc. ICSP*, 2010, pp. 502 - 505.

[7] R. Tucker, “Voice activity detection using a periodicity measure,” *Proc. IEEE*, vol. 139, pp. 377–380, Aug. 1992.

[8] Yanmeng Guo; Kai Li; Qiang Fu; Yonghong Yan, "A two-microphone based voice activity detection for distant-talking speech in wide range of direction of arrival," *IEEE ICASSP*, 2012, vol., no., pp.4901,4904, 25-30 March 2012.

[9] Rubio, J.E.; Ishizuka, Kentaro; Sawada, H.; Araki, S.; Nakatani, T.; Fujimoto, M., "Two-Microphone Voice Activity Detection Based on the Homogeneity of the Direction of Arrival Estimates," *IEEE ICASSP* 2007, vol.4, no., pp.IV-385,IV-388, 15-20 April 2007

[10] E. Nemer, R. Goubran, and S. Mahmoud, “Robust voice activity detection using higher-order statistics in the LPC residual domain,” *IEEE Trans. Speech Audio Process*., vol. 9, no. 3, pp. 217–231, Mar. 2001.

[11] E. Nemer, “Speech Analysis and Quality Enhancement Using Higher-Order Cumulants,” *Ph.D. dissertation*, Carleton Univ., Ottawa, ON, Canada, 1999.

[12] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection", *IEEE Signal Processing Letters*., vol. 6, pp.1 -3 1999.

[13] Yong Duk Cho; Kondoz, A., "Analysis and improvement of a statistical model-based voice activity detector," *Signal Processing Letters, IEEE* , vol.8, no.10, pp.276,278, Oct. 2001

[14] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp.1109 -1121 1984

[15] A. Guerin. “Two-Sensor Noise Reduction System: Applications for Hands-Free Car Kit”. *EURASIP Journal on Applied Signal Processing* 2003:11, 1125–1134.