Noise Robust Voice Activity Detection Using Features Extracted From the Time-Domain Autocorrelation Function

Houman Ghaemmaghami, Brendan Baker, Robbie Vogt, Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia

{houman.ghaemmaghami,bj.baker,r.vogt,s.sridharan}@qut.edu.au

Abstract

This paper presents a method of voice activity detection (VAD) for high noise scenarios, using a noise robust *voiced* speech detection feature. The developed method is based on the fusion of two systems. The first system utilises the maximum peak of the normalised time-domain autocorrelation function (MaxPeak). The second system uses a novel combination of cross-correlation and zero-crossing rate of the normalised autocorrelation to approximate a measure of signal pitch and periodicity (CrossCorr) that is hypothesised to be noise robust. The score outputs by the two systems are then merged using weighted sum fusion to create the proposed autocorrelation zero-crossing rate (AZR) VAD. Accuracy of AZR was compared to state-of-the-art and standardised VAD methods and was shown to outperform the best performing system with an average relative improvement of 24.8% in half-total error rate (HTER) on the QUT-NOISE-TIMIT database created using real recordings from high-noise environments.

Index Terms: voice activity detection, high noise, autocorrelation, zero-crossing rate, time-domain analysis

1. Introduction

Voice activity detection (VAD) is an essential technique in speech processing. It is commonly utilised in automatic speech recognition, speaker recognition, speaker diarisation, and speech enhancement [1]. This has led to the proposal of a variety of VAD algorithms, however, the need for noise robust and efficient speech processing methods has increased, and thus, so has the need for high precision VAD algorithms that operate under extremely noisy conditions - scenarios with a signal-to-noise ratio (SNR) of <5 dB. This has been the motivation behind the design of the autocorrelation zero-crossing rate (AZR) system presented in this paper.

VAD algorithms typically consist of a feature extraction stage followed by a classification/segmentation process. Some of the features utilised include: energy, zero crossing rate, cepstral coefficients [1], higher order spectra (HOS) [2], LPC analysis [3], autocorrelation [4], spectral divergence [4], harmonic features [5], and spectral pattern [6]. Most VAD methods utilise a combination of features. The classification methods employed in recent VAD algorithms include: Gaussian mixture models (GMM) [7], support vector machine (SVM) [8], and Gaussian likelihood ratio test (LRT) [9].

Analysis of available VAD methods reveals that such algorithms are commonly more successful in the detection of voiced speech segments rather than unvoiced speech. This is expected, as the statistical nature of unvoiced speech is more random and typically more similar to background noise [4-6]. It is the quasi-periodic nature of voiced speech that makes it different to unvoiced speech and thus more robust to background noise [10]. For this reason, the AZR VAD system is implemented to focus on the detection of voiced speech segments. To do this, two features for conducting VAD, and specifically *voiced* speech detection, in high noise scenarios are proposed. Each feature is implemented as an autonomous system that carries out an independent measure of *voicing*. The first system utilises the maximum peak value of the normalised time-domain autocorrelation. This feature is titled the MaxPeak score. The second system utilises a combination of the zero-crossing rate and cross-correlation of the time-domain autocorrelation to provide an approximate measure of signal pitch and periodicity. This feature is titled the CrossCorr score. The raw output scores from the two systems are fused using a weighted sum fusion. The fused scores are then thresholded and smoothed to attain final speech/non-speech decisions. These decisions are then utilised to evaluate the proposed method with respect to reference VAD data.

Section 2 provides a description of the proposed AZR VAD through describing the MaxPeak and CrossCorr algorithms in Sections 2.1 and 2.2, respectively. The fusion of the two algorithms and the smoothing process used to estimate unvoiced speech frames is explained in Section 2.3. Section 3 provides a brief description of the QUT-NOISE-TIMIT database [11] utilised for evaluation, and Section 4 describes the experiments conducted and results obtained. These results are compared to the performance of the ITU-T G.729 Annex B [12], advanced front-end (AFE) ETSI [13], long term spectral divergence (LTSD) [14], and Sohn's likelihood ratio test (LRT) VAD [9] systems. It is shown that the developed system provides greater accuracy, than the baseline methods, across all tested noise levels and scenarios.

2. AZR VAD

The method proposed in this paper is designed to obtain voiced speech segments. This method is titled AZR and is implemented as a fusion of two independent systems, MaxPeak and CrossCorr. Each of the two systems utilise features extracted from the autocorrelation function to produce voiced speech scores. Unvoiced speech segments are estimated using smoothing in the AZR VAD.

Voiced speech is quasi-periodic and has a pitch. The pitch frequency for an adult typically ranges from 50 to 500 Hz [5]. This corresponds to pitch periods of 2 to 20 ms, hence, the time-domain autocorrelation is calculated using a frame size of (> 20 ms). A frame size of 50 ms, with no overlapping frames was chosen for both the MaxPeak and CrossCorr algorithms. Experiments with overlapping frames demonstrated no real benefit over no-overlap, thus a non-overlapping implementation was chosen to reduce computation time.

2.1. MaxPeak Algorithm

One method of classifying voiced speech is using the maximum peak of the normalised autocorrelation within the lag range corresponding to the expected pitch periods of voiced speech (2 to 20ms) [4, 5]. The MaxPeak system

follows this process and outputs a score vector containing the values of this peak per analysed frame.

The MaxPeak system segments an input signal, $s[i]$ (where i represents sample number), into k, 50 ms frames, such that,

$$s[i] = \{s_1[i]|s_2[i]| \dots |s_k[i]\} \quad (1)$$

a DC removal and pre-emphasis is then applied to each frame,

$$x_k[i] = (s_k[i] - \mu_k) - \alpha(s_k[i-1] - \mu_k) \quad (2)$$

where μ_k is the mean of $s_k[i]$ (the k^{th} frame of $s[i]$) and $x_k[i]$ is the pre-emphasised and DC removed version of $s_k[i]$.

The pre-emphasis constant α, in (2), is typically chosen between 0.9 and 1. In this case α is set to 0.96 and the pre-emphasis is conducted to correct the roll-off in the spectrum of voiced speech that is caused by radiation from the mouth and the voiced excitation source [10].

After pre-emphasis, the normalised time-domain autocorrelation, $R_k[z]$, at lags corresponding to pitch periods of 2 to 20 ms, is calculated for $x_k[i]$,

$$R_k[z] = \frac{\sum_{i=1}^{n-z} x_k[i]x_k[i+z]}{\sum_{i=1}^{n} x_k^2[i]} \quad (3)$$

where z is the autocorrelation lag and n is the number of samples in $x_k[i]$, hence, the MaxPeak score ($M[k]$), for the k^{th} frame, $x_k[i]$, is calculated as,

$$M[k] = \max(R_k[z]) \quad (4)$$

it is expected that voiced speech would produce a higher maximum peak value than unvoiced or silence/noise frames.

2.2. CrossCorr Algorithm

The MaxPeak feature alone cannot serve as a noise robust VAD feature. It has been shown that the maximum peak value of the normalised autocorrelation reduces as SNR decreases [4], making it more difficult to distinguish between voiced and unvoiced/noise frames. For this reason a novel feature, titled CrossCorr, was developed to conduct noise robust *voiced* speech detection.

The quasi-periodic nature of voiced speech causes the autocorrelation function of voiced frames to be approximately periodic within the 2 to 20 ms lag range. It is hypothesised that this "periodicity" is more robust to noise than the MaxPeak feature. In the case of high noise scenarios, it can be observed that while the peak value of the normalised autocorrelation drops, the "periodicity" is preserved. A novel feature, CrossCorr, is proposed to measure this "periodicity" and thus distinguish between voiced and unvoiced/noise frames at low SNR.

The CrossCorr algorithm follows the same process as that in section 2.1 to obtain the normalised autocorrelation function $R_k[z]$, however, pre-emphasis filtering is not carried out in order to preserve the original pitch of the input signal.

The pitch of the signal can be approximated by calculating the zero-crossing rate of the autocorrelation. The CrossCorr system utilises the zero-crossing rate of the autocorrelation within the 2 to 20 ms lag range to selectively analyse signals with approximate pitch frequencies corresponding to a pitch range of 50 to 500 Hz. This acts as an initial screening process and is done to ensure further analysis is only carried out on signals that are deemed as potential voiced speech segments.

A signal is defined as periodic if it repeats its values every period. If a periodic signal is segmented into its periods, the segments will display perfect correlation with one another. The CrossCorr method utilises this definition loosely,

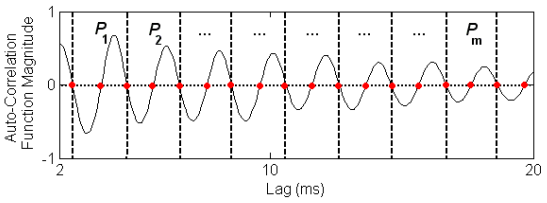

Figure 1: *$R_k[z]$ of a voiced speech frame for 2 to 20 ms lag. The CrossCorr algorithm initially performs a count of the marked zero-crossing points to estimate the pitch and perform a cross-correlation similarity check of adjacent P_y segments for $y = 1, 2, \dots, m$.*

suggesting that a signal is "more periodic" if a higher correlation exists between its "periods", and "less periodic" if the opposite is true. The CrossCorr method first segments the 2 to 20 ms lag range of the autocorrelation ($R_k[z]$), for a given frame, into its "periods". To do this, it is assumed that a "period" is observed every two zero-crossing points. Each segment is then cross-correlated with its posterior segment and the summation of the maximum values obtained from each cross-correlation is recorded as the CrossCorr score, $C[k]$, for the analysed frame. It must be noted that the cross-correlated "periods" will not have equal lengths in most cases and are thus zero-padded to adjust segment lengths for the cross-correlation process.

Figure 1 shows that once the zero-crossing rate of $R_k[z]$, which indicates the approximate pitch of the analysed frame, satisfies the specified upper and lower limits (it is within the approximate 50 to 500 Hz pitch period) the CrossCorr, $C[k]$, feature can then be calculated using (5) and (6),

$$\hat{R}_y[z'] = \sum_{j=1}^{n'-z'} P_y[j]P_{y+1}[j+z'] \quad (5)$$

therefore,

$$C[k] = \sum_{y=1}^{m-1} \max(\hat{R}_y[z']) \quad (6)$$

where P_y specifies an assumed "period" of $R_k[z]$, and $\hat{R}_y[z']$ is the cross-correlation function between P_y and its posterior "period". This adjacent cross-correlation is used to ensure maximum CrossCorr score value in cases such as that in Figure 1, where the magnitude of the autocorrelation decreases with lag increase.

The CrossCorr feature, $C[k]$, is thus an indicator of "periodicity" within a desired pitch range and is therefore robust to both random and quasi-periodic noise. Figures 2 displays the projection of the "periods" of $R_k[z]$for voiced and unvoiced/noise frames onto one another, which indicates the correlation of these "periods" in each case.

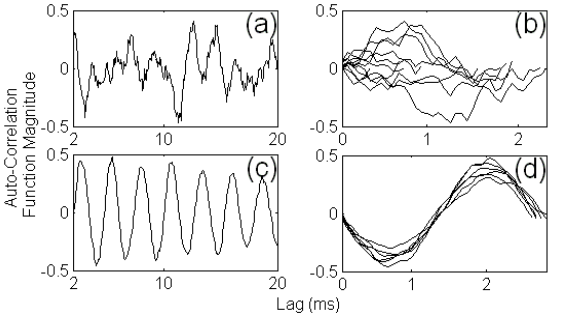

Figure 2: *(a) $R_k[z]$ of a noise frame (2 to 20 ms lag) from a noisy speech file at SNR=(-5)dB. (b) Similarity check of P_y "periods" every second zero-crossing indicates low correlation. (c) $R_k[z]$ of a voiced speech frame at SNR=(-5)dB. (d) Similarity check of P_y "periods" indicates high correlation.*

2.3. Fusion and Smoothing

A normalisation process was carried out prior to the weighted sum fusion of the output scores. This was conducted to achieve a wider score range for the MaxPeak system, efficiently employ the fusion, and obtain the AZR scores. To do this, the MaxPeak and CrossCorr scores were independently calculated, and the MaxPeak scores, which initially covered an expected range of $(0 < M[k] < 1)$ were modified according to,

$$M'[k] = -\log(1 - M[k]) \quad (7)$$

scores were then divided into two independent sets, location-1 and 2, as described in Section 3. The medians of the speech score distributions for each location set, of each system, were calculated and utilised to normalise the scores for the independent location set of that system to obtain unbiased results. This score normalisation was followed by the weighted sum fusion. Experiments were conducted for various weighted summations. It was observed that maximum accuracy was achieved using an equal weighted sum to obtain AZR scores:

$$AZR[k] = M'[k] + C[k] \quad (8)$$

The AZR method conducts *voiced* speech detection. For this reason the unvoiced frames must be estimated to complete the VAD process. Based on the study in [4], voiced speech is typically preceded by 300ms and followed by 500ms of unvoiced speech, hence, a moving average smoothing filter, with a length of 1 second, was employed to smooth the raw AZR scores and compensate for the missed unvoiced frames. The smoothed scores were then thresholded and segmented to complete the AZR VAD.

3. QUT-NOISE-TIMIT Corpus

The QUT-NOISE-TIMIT database [11], consisting of 600 hours of noisy speech was produced and utilised for testing. The files were created at set lengths of 60 and 120 seconds at various SNR, using clean speech from the TIMIT database and realistic noise recordings from the QUT-NOISE corpus. A total of 24,000 files were created and used for testing.

The noise recordings were collected from 10 independent real-noise locations, with each recording having a length of 30 minutes. The clean speech files were then added, at random, to random selections of the real-noise recordings to produce noisy speech files at set SNR levels and recording length for each scenario location. The files had a sampling rate of 16,000 Hz and were grouped based on noise type and location.

It can be seen from Table 1 that the database consists of 5 distinct real-life recorded noise scenarios, with each divided into two equal sized sets based on recording location or noise type. Each location is then divided into 3 equal sized noise-level sets: low noise (SNR=10 or 15 dB), medium noise (0 or 5 dB), and high noise (-5 or -10 dB). Each set contains an equal number of files. For example, CAFE scenario, at location "Food court", contains 2,400 noisy speech files, with 1,200 of the files having a length of 60 seconds and the remaining 1,200 files at 120 seconds. In addition, the set contains 800 files at each noise level.

4. Experiment

The proposed AZR VAD was evaluated against state-of-the-art methods, long term spectral divergence (LTSD) [14] and Sohn's likelihood ratio test (LRT) VAD [9], and standardised VAD systems, advanced front-end (AFE) ETSI [13] and the ITU-T G.729 Annex B [14].

Table 1. *QUT-NOISE-TIMIT scenarios and locations.*

Scenarios	Location 1	Location 2
CAFE	Cafe	Food court
CAR	Window down	Window up
HOME	Kitchen	Living room
REVERB	Car park	Pool
STREET	City	Suburb

4.1. Evaluation

Errors were calculated as percentages of time. The metrics utilised included, false alarm rate (FAR), miss rate (MR), and half-total error rate (HTER). HTER was calculated as the average of FAR and MR. The errors were obtained with respect to reference speech/non-speech boundaries.

$$\%\ \text{FAR} = \left(\frac{nonspeech\ samples\ detected\ as\ speech}{Total\ number\ of\ nonspeech\ samples}\right) \times 100$$

$$\%\ \text{MR} = \left(\frac{speech\ samples\ detected\ as\ nonspeech}{Total\ number\ of\ speech\ samples}\right) \times 100$$

4.2. Training and Testing

To obtain unbiased test results over the entire corpus, thresholds for the location-1 scenarios were generated using systems trained on the location-2 data, and vice versa. That is, a 2-fold cross-validation approach was employed to minimise HTER with folds defined according to noise location. This was employed for the training and testing of the developed and baseline methods, except G.729-B and ETSI systems, as their parameters are fixed according to their standard specification.

4.3. Results

Figure 3 displays the HTER percentages, per noise level and scenario, as bar graphs for each VAD system. Each bar consists of a light and dark shade of colour. The dark shade represents the proportion of the HTER that is caused by the MR, while the lighter shade indicates the contribution of the FAR to the HTER. The exact error percentages obtained at each noise level, for the systems, can be found in Table 2.

From Figure 3, high FAR error proportions are observed for all systems in the CAFE and HOME scenarios. This can be attributed to the presence of background speech in the CAFE and HOME scenarios. This background speech may be falsely considered as target speech, especially at negative SNR, thus causing the FAR errors.

High MR error proportions and HTER, are observed for all systems in the REVERB scenario, with the exception of ETSI displaying a high FAR proportion. This may be due to reverb corrupting the speech signal. It is, however, apparent that AZR is less affected by reverb with respect to baseline methods.

From Table 2, it can be concluded that AZR outperforms the baseline methods over the entire database. The AZR VAD achieves 33.6%, 28.5%, and 12.2% relative improvements in HTER over the best performing baseline VAD (LTSD) for the low, medium, and high-noise scenarios, respectively.

5. Conclusion

In this paper, a noise robust VAD was proposed based on, only the detection of voiced speech. It was suggested that the quasi-periodic nature of voiced speech, captured by the autocorrelation function, can be utilised as a noise robust feature while unvoiced speech can be approximated using smoothing. This was achieved through the implementation of a

novel feature (CrossCorr), which employs the cross-correlation and zero-crossing rate of the normalised autocorrelation to approximate signal pitch and conduct a periodicity measure. The pitch approximation ensures that the analysed signal is within a reasonable pitch range, with reference to the pitch of a typical adult, and the periodicity measure indicates the level of periodicity of the signal.

The maximum peak of the normalised autocorrelation (MaxPeak) was utilised to aid the *voiced* speech detection. The two features were fused using summation to obtain the proposed AZR VAD. AZR was evaluated against four baseline systems using the QUT-NOISE-TIMIT corpus. It was shown that AZR outperforms the best performing baseline system (LTSD) with an average relative improvement of 24.8%.

Table 2. *Overall %FAR, %MR, and %HTER for the AZR VAD and baseline systems at each tested noise level.*

VAD Systems	Low Noise (SNR=10 or 15dB)			Medium Noise (SNR=0 or 5dB)			High Noise (SNR=-10 or -5dB)		
	% FAR	% MR	% HTER	% FAR	% MR	% HTER	% FAR	% MR	% HTER
AZR	15.6	6.6	**11.1**	20.5	12.1	**16.3**	31.9	25.5	**28.7**
LTSD	20.7	12.8	**16.7**	26.2	19.5	**22.8**	28.6	36.8	**32.7**
Sohn's (LRT)	24.6	20.4	**22.5**	33.5	28.9	**31.2**	56.5	25.0	**40.8**
G.729-B	33.7	18.8	**26.2**	34.2	31.5	**32.9**	35.0	50.2	**42.6**
ETSI	68.3	0.2	**34.2**	66.8	2.2	**34.5**	65.1	13.6	**39.4**

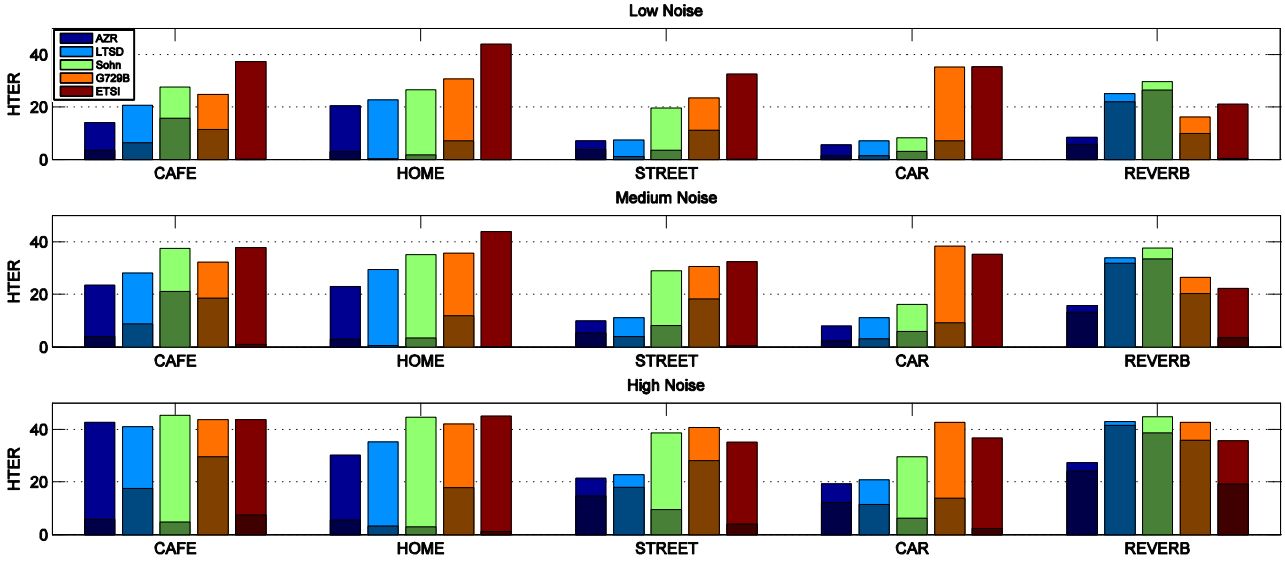

Figure 3: *%HTER performance of AZR VAD and baseline methods for each noise scenario at three tested noise levels. The dark shading of each bar represents the %MR and lighter shade displays the %FAR of the overall %HTER.*

6. References

[1] Haigh, J. A., Mason, J. S., "Robust voice activity detection using cepstral features," in TENCON '93. Conference on Computer, Communication, Control and Power Engineering Proceedings.1993 IEEE Region 10, 1993, pp. 321-324 vol.3.

[2] Rangoussi, M., Delopoulos, A., Tsatsanis, M., "On the use of higher-order statistics for robust endpoint detection of speech," in Higher-Order Statistics, 1993., IEEE Signal Processing Workshop on, 1993, pp. 56-60.

[3] Nemer, E., Goubran, R., Mahmoud, S., "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.

[4] Kristjansson, T., Deligne, S., Olsen, P., "Voicing Features for Robust Speech Detection," Interspeech-Eurospeech, submitted, 2005, Lisboa, Portugal.

[5] Dou, H., Bao, C., Li, R., "A voice activity detection using cyclic statistics based on sinusoidal speech model," International Conference on Signal Processing, 2008. ICSP 2008. 9th, vol., no., pp.1239-1242, 26-29 Oct. 2008.

[6] Fisher, E., Tabrikian, J., Dubnov, S., "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," IEEE Transactions on Audio, Speech, and Language Processing, vol.14, no.2, pp. 502- 510, March 2006.

[7] Abdulla, W.H., Guan, Z., Sou, H. C., "Noise robust speech activity detection," International Symposium on Signal Processing and Information Technology (ISSPIT), IEEE 2009, vol., no., pp.473-477, 14-17 Dec. 2009.

[8] Ramirez, J., Yelamos, P., Gorriz, J. M., "SVM-based speech endpoint detection using contextual speech features," Electronics Letters , vol.42, no.7, pp. 426-428, 30 Mar. 2006.

[9] Sohn, J., Kim, N. S., Sung, W., "A statistical model-based voice activity detection," Signal Processing Letters, IEEE , vol.6, no.1, pp.1-3, Jan 1999.

[10] Eom, K. B., Chellappa, R., "Classification of voiced and unvoiced speech by hierarchical stochastic modeling," International Conference on Pattern Recognition, 1994. Vol. 3 - Conference C: Signal Processing, Proceedings of the 12th IAPR , vol., no., pp.20-24 vol.3, 9-13 Oct 1994.

[11] Dean, D., Vogt, R., Mason, M., Sridharan, S., "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," Submitted to Interspeech 2010.

[12] ITU-T Recommendation G.729 Annex B, "A silence compression scheme for G.729 optimized for terminals conforming to recommendation v.70," 1996.

[13] Li, J. Y., Liu, B., Wang R. H., Dai, L. R., "A complexity reduction of ETSI advanced front-end for DSR," International Conference on Acoustics, Speech, and Signal Processing, Proceedings. (ICASSP '04). IEEE , vol.1, no., pp. I-61-4 vol.1, 17-21 May 2004.

[14] Ramirez, J., Segura, J.C., Benitez, C., de la Torre, A., "Voice activity detection with noise reduction and long term spectral divergence estimation," Int. Conference on Acoustics, Speech, and Signal Processing, Proceedings. Proceedings. IEEE, vol.2, no., pp. ii-1093-6 vol.2, 17-21 May 2004.