A Robust Endpoint Detection Algorithm Based on Identification of the Noise Nature

Denilson C. Silva

Program of Electrical Engineering - COPPE
Signal Processing Laboratory
Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
denilson@lps.ufrj.br

Abstract

The endpoint detection of speech is still a big problem in situations of speech recognition in noisy environments. While traditional methods concentrate on finding speech in noise, the proposed technique is based on noise identification through HMMs, associated with both *SNR* and euclidean distance of the log-energy calculated on a frame-by-frame basis. Computer experiments confirm that the proposed algorithm gives rise to a considerable improvement on the precision of endpoint detection, specially in severely adverse conditions where the *SNR* is very low.

1. Introduction

In many applications of speech signal processing, determination of the endpoints of utterances is necessary. The traditional methods of endpoints detection based on both energy and zero crossing rate work very well with clean speech [1]. When we have utterances with fricatives, for instance, the endpoint detection might become complicated if the delimitation process happens in a noisy environment.

Several works have been seeking to solve the subject of the endpoint detection in noisy environments [2, 3], but the obtained results are extremely sensitive to the signal-to-noise ratio (*SNR*).

In this article a method is proposed for endpoint detection in utterance for speech recognition based on the principle of identification of the noise nature that contaminates the signal, through a classification on each frame using hidden Markov models (HMMs), delimiting the intervals with speech starting from the identification of frames with noise only. The *SNR* and the euclidean distance of the log-energy of each frame are also used to accomplish a refinement in the detection. The proposed method results in a significant precision improvement, particularly for very adverse conditions, as computer experiments show.

This article is organized as follows. In Section 2, the proposed endpoint detection algorithm is introduced. In Section 3, the database is described. In Section 4, results of the accomplished tests as well as comparisons with the traditional method based on both energy and zero crossing rate are shown. Finally, in Section 5, we present the conclusions.

2. Proposed method

The proposed endpoints detection method is accomplished through three processes of decision beginning with a frame-by-frame analysis: the identification of the noise nature [4] [5], *SNR* and the euclidean distance of the log-energy (Figure 1). We define that a frame with positive flag $(+1)$ contains noise only, while a frame with negative flag (-1) contains both speech and noise.

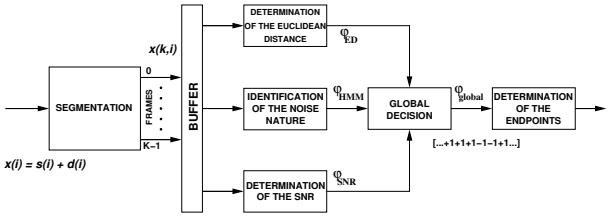

Figure 1: *Proposed method.*

2.1. Identification of the noise nature

Consider the noisy signal $x(i)$ a composition of the clean speech $s(i)$ with the additive noise $d(i)$, that is, $x(i) = s(i) + d(i)$.

The signal $x(i)$ is segmented into K frames with N size and overlap of L samples:

$$x(k,i) = x(k(N-L)+i) \tag{1}$$

where $0 \leq k \leq K-1$ and $0 \leq i \leq N-1$.

The process of identification of the noise nature is accomplished through a classification of the frames of the noisy signal ($x(k,i)$), using HMMs, among the types of noise involved in the training. The settings used on tests were 18 parameters per subframes of 23 ms, 50% of overlap, Hamming windowing, left-right HMM with skip allowed, four states, four models and 128 centroids. The stages of the identification are explained below.

2.1.1. Extraction of parameters

The extracted parameters on each subframe were: one of spectral entropy (as described in [3]), one of zero crossing rate and 16 of log-energy, described in Table 1.

The pth log-energy parameter, $LogE(k,p)$, is defined as the energy contained in the pth subband for frame k.

2.1.2. Criterion of decision

The noisy utterances are segmented in frames of 46 ms and overlap of 80%. The frames are identified by HMM, in agreement with trained noise models. The results are registered as

Table 1: Algorithm for extraction of log-energy parameters.

1. Consider the noisy signal spectrum, $\Gamma(k,j)$, calculated by FFT.
2. Calculate the log-spectrum ($\Psi(k,j)$):
$\Psi(k,j) = 10\log 10\left(\frac{\Gamma(k,j)}{f_s m}\right)$, where $0 < j < F-1$, F is the number of frequency components, f_s is the sampling frequency and m is the number of points used in the FFT;
3. Estimate the log-spectral envelope ($\Phi(k,j)$):
a. Initialization:
$\Phi(k,0) = \Psi(k,0)$
$\Phi(k,F-1) = \Psi(k,F-1)$
b. Iteration:
for $j = 1$ to $F-2$ **do**
  if $\Psi(k,j) > \Psi(k,j-1)$ and
  $\Psi(k,j) > \Psi(k,j+1)$ **then**
    $\Phi(k,j) = \Psi(k,j)$
  else
    $\Phi(k,j) = 0$
  end if
end for
c. Interpolate the values between each pair of adjacent non-null values by Newton's divided difference method [6].
5. Divide the log-spectral envelope into P subbands with F' frequencies in each subband.
6. Extract P log-energy parameters ($LogE(k,p)$).
$LogE(k,p) = \sum_{j=0}^{F'-1} \Phi(k,(pF'+j))$, where $0 < p < P-1$

noise frame. In the following, a count is accomplished to verify which noise type received the largest number of classifications in all signal (*noise signal*).

$\varphi_{HMM}(k)$ is the flag attributed to the *k*th frame.

$$\varphi_{HMM}(k) \quad = \quad \begin{cases} -1, & \text{if } \textit{noise frame} \neq \textit{noise signal} \\ +1, & \text{if } \textit{noise frame} = \textit{noise signal} \end{cases}$$

2.2. Determination of the *SNR*

The *SNR* of each frame is calculated in accordance with the algorithm in Table 2 and a flag $\varphi_{SNR}(k)$ is assigned.

$th1$ received empirically the value of 1.25, if $\xi < 10$dB, and 2.50, if $\xi \geq 10$dB. M and Q were attributed values 15 and 5, respectively.

ξ is the estimated *SNR* for all signal, calculated as:

$$\xi = 10\log 10\left(\frac{\sum_{k=0}^{K'-1}\hat{\sigma}_x^2(k) - \sum_{k=0}^{K'-1}\hat{\sigma}_d^2(k)}{\sum_{k=0}^{K'-1}\hat{\sigma}_d^2(k)}\right) \quad (2)$$

where both $\hat{\sigma}_x^2(k)$ and $\hat{\sigma}_d^2(k)$ are calculated such as $\sigma_x^2(k)$ and $\sigma_d^2(k)$, from algorithm in Table 2, without overlap of the K' frames, in other words, $x(k,i) = x(kN+i)$, for $0 \leq k \leq K'-1$.

2.3. Determination of the euclidean distance

The euclidean distance is evaluated based on vectors obtained from the log-energy parameters of $x(i)$. The flag assignment algorithm can be found in Table 3.

Table 2: Algorithm for determination of the flags by *SNR*.

1. Calculate the variance of noisy signal $x(k,i)$ ($\sigma_x^2(k)$).
$\sigma_x^2(k) = \frac{1}{N}\sum_{i=0}^{N-1}[x(k,i) - \mu(k)]^2$
a. Here $\mu(k) = \frac{1}{N}\sum_{i=0}^{N-1} x(k,i)$
2. Estimate the variance of noise using a smoothing filter [7] ($\sigma_d^2(k)$).
a. Considering M frames inside the initial interval of the utterance, without the presence of the speech signal $s(i)$, calculate the relative *SNR* ($\Xi(k)$).
$\Xi(k) = \frac{\sigma_x^2(k)}{\frac{1}{M}\sum_{n=0}^{M-1}\sigma_x^2(n)}$
b. Calculate the $\alpha(k)$ parameter.
$\alpha(k) = 1 - min(1, \Xi(k)^{-Q})$
c. Estimate noise.
$\sigma_d^2(k) = \alpha(k)\sigma_d^2(k-1) + (1-\alpha(k))\sigma_x^2(k)$
3. Define flag.

$$\varphi_{SNR}(k) \quad = \quad \begin{cases} -1, & \text{se } \sigma_x^2(k)/\sigma_d^2(k) \geq th1 \\ +1, & \text{se } \sigma_x^2(k)/\sigma_d^2(k) < th1 \end{cases}$$

$\varphi_{ED}(k)$ is the flag attributed to *k*th frame by euclidean distance criterion. $th2$ received empirically the value of 1.4.

2.4. Determination of the endpoints

After all frames receive three flags, they are labeled again, receiving one flag only, ($\varphi_{global}(k)$):

if $\varphi_{HMM}(k) = \varphi_{SNR}(k) = \varphi_{ED}(k) = +1$ **then**
  $\varphi_{global}(k) = +1$
else
  $\varphi_{global}(k) = -1$
end if

The global flags are analyzed starting from frame 0 in direction to the last, looking for the first sequence of 15 frames with negative flags, where the first frame of the sequence characterizes the ending of a possible initial interval containing only noise. Starting from the last frame in direction to the frame 0, a sequence of 15 frames with negative flag is also sought, where the last frame of the sequence characterizes the beginning of a possible final interval containing only noise.

3. Database

The utterances used in this article were collected of the database described in [8], where we have 10 command and control isolated words. The sampling rate is 11025 Hz.

The noise database was collected from [9], with original sampling rate of 19980 Hz, and resampling to 11025 Hz. Four noise types were selected: WHITE, PINK, VOLVO (car interior) and BABBLE.

The noisy speech was formed through the addition of selected noise to clean speech with *SNR* from 0 to 20dB.

4. Experimental Results

The performance could be appraised comparing the results obtained in detection with obtained reference values of manual

Table 3: Algorithm for determination of the flags by euclidean distance.

1. Calculate the euclidean distance $ED_x(k)$, on each frame, between the log-energy vectors of signal $x(i)$ and a reference vector.
a. Calculate the reference vector, $LogE_{ref}(p)$
$LogE_{ref}(p) = \frac{1}{M}\sum_{k=0}^{M-1} LogE(k,p)$
b. Calculate the distance.
$ED_x(k) = \sqrt{\sum_{p=0}^{P-1}(LogE(k,p) - LogE_{ref}(p))^2}$
2. Estimate ED_d.
$ED_d = \max_{0\leq k\leq M-1}\{ED_x(k)\}$
3. Define flag.

$$\varphi_{ED}(k) \quad = \quad \begin{cases} -1, & \text{if } ED_x(k)/ED_d \geq th2 \\ +1, & \text{if } ED_x(k)/ED_d < th2 \end{cases}$$

Table 4: Table containing average percents of error reduction rate

	BABBLE	PINK	VOLVO	WHITE
Beginning	6.82%	12.23%	−2.65%	12.05%
Ending	20.81%	26.24%	0.70%	24.88%

clipping.

Initially it was made a training of discrete HMMs of [8], with 500 segments of 100 ms for each of the four types of noise and 50% of overlap. An experiment was undertaken with 121 corrupted utterances, setting both the *SNR* and the noise, observing the percent of mistake in detection along the several values of *SNR* both at beginning (ε_b) and at ending (ε_e), where B and E are, respectively, the beginning and the ending by manual clipping, $\mathcal{E}_b$ and $\mathcal{E}_e$ are, respectively, the detected points by the proposed system.

$$\varepsilon_b = \frac{|B - \mathcal{E}_b|}{E - B} \times 100\% \tag{3}$$

$$\varepsilon_e = \frac{|E - \mathcal{E}_e|}{E - B} \times 100\% \tag{4}$$

The average reduction of the error rate in detection, comparatively to the method described in [1], is shown in Table 4. We can see that there was a considerable error reduction in every case, except for the car interior noise, where no significant change was observed.

The results for each type of noise of the traditional [1] and proposed technique can be found in Tables 5, 6, 7 and 8. It is also noticed that the best performance of the introduced detector happens in low-*SNR* zone.

Figure 2 exemplifies the endpoint detection on low-energy sounds. We can see that the proposed method kept all useful information signal.

5. Conclusions

In this paper a robust endpoint detection method was proposed based on identification of the noise nature using

Table 5: Percentual error of the endpoint detection in babble noise environment

SNR	Traditional		Proposed	
	beginning	ending	beginning	ending
0dB	40.0%	80.67%	11.96%	20.55%
5dB	15.83%	39.41%	9.38%	10.78%
10dB	8.95%	19.11%	7.28%	8.66%
15dB	6.54%	10.85%	6.87%	6.33%
20dB	5.34%	6.04%	7.04%	5.71%

Table 6: Percentual error of the endpoint detection in pink noise environment

SNR	Traditional		Proposed	
	beginning	ending	beginning	ending
0dB	68.52%	114.26%	15.20%	36.55%
5dB	19.07%	53.69%	11.87%	20.35%
10dB	14.04%	34.13%	11.18%	17.32%
15dB	9.79%	14.66%	10.08%	9.95%
20dB	7.10%	6.63%	9.03%	8.04%

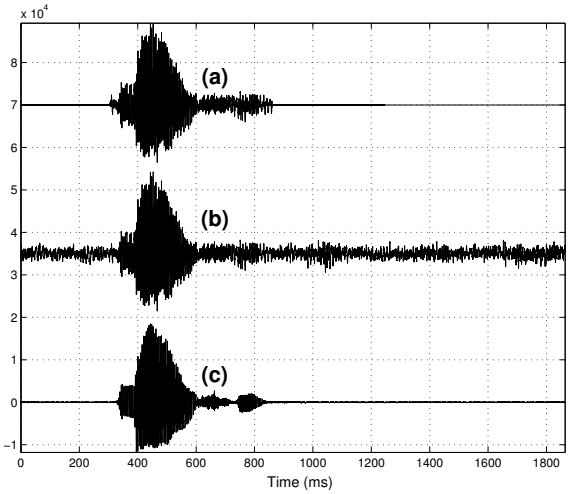

Figure 2: Endpoint detection based on identification of the noise nature. (a) Endpoint detection by the proposed method. (b) Corrupted signal by babble noise at 5dB *SNR* and (c) Clean speech referring to "baixo" (which means "down") utterance.

HMMs, associated to *SNR* and to euclidean distance of log-energy parameters, that were used to adjust the detection in segments with low-energy sounds. With this method we try to detect the frames that contain only noise and to separate them from the useful information of the signal. Computer experiments with 121 utterances corrupted by four different noise types and with varying levels of *SNR* from 0 to 20dB were presented. It can be seen from the results that in severe adverse conditions, when the *SNR* is very low, that the proposed method offers a considerably more precise detection of endpoints as compared to the traditional technique. The proposal of an algorithm that, based on the estimation of *SNR* and type of noise, can decide

Table 7: Percentual error of the endpoint detection inside car

SNR	Traditional		Proposed	
	beginning	ending	beginning	ending
0dB	9.67%	15.02%	7.52%	5.64%
5dB	7.43%	10.86%	8.22%	6.08%
10dB	7.55%	6.71%	9.38%	6.16%
15dB	8.40%	5.04%	12.66%	8.24%
20dB	7.09%	4.53%	15.61%	12.53%

Table 8: Percentual error of the endpoint detection in white noise environment

SNR	Traditional		Proposed	
	beginning	ending	beginning	ending
0dB	62.44%	105.94%	13.04%	29.29%
5dB	18.31%	50.79%	11.49%	21.94%
10dB	13.68%	33.22%	9.87%	16.44%
15dB	9.61%	15.19%	8.05%	9.25%
20dB	6.93%	6.57%	8.29%	10.37%

which is the most suitable technique for end point detection and recognition in adverse conditions is subject of ongoing research.

6. References

[1] Teruszkin, R., Consort, T. A. and Resende Jr., F. G. V., "Endpoint detection analysis for an implementation of a speech recognition system applied to robot control", In Proc. SAWCAS, November, 2001.

[2] Bou-Ghazale, S. E. and Assaleh, K., "A robust endpoint detection of speech for noisy environment with application to automatic speech recognition", In IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.4, pp. 3808-3811, May, 2002.

[3] Jia-Lin Shen, Jeih-Weih Hung and Lin-Shan Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments", In International Conference on Spoken Language Processing, Sydney, November, 1998.

[4] Silva, D. C. and Resende Jr., F. G. V., "Identification of the Noise Nature based on HMMs" (in Portuguese), In Simpósio Brasileiro de Telecomunicações, September, 2004.

[5] Silva, D. C., "Identification of the Noise Nature with Application in Robust Speech Recognition" (in Portuguese), Masters Thesis, Federal University of Rio de Janeiro, February, 2005.

[6] Hildebrand, F. B., Introduction to Numerical Analysis, McGraw-Hill, New York, Second Edition, 1974.

[7] Lin, L., Holmes, W. H. and Ambikairajah, E., "Subband noise estimation for enhancement using a perceptual Wiener filter", In IEEE International Conference on Acoustics, Speech, and Signal Processing, vol.1, pp.80-83, April, 2003.

[8] Teruszkin, R., Resende Jr., F. G. V., Villas-Boas, S. B. and Lizarralde, F., "Object-oriented speech recognition library applied to robot control" (in Portuguese), In Congresso Brasileiro de Automática, September, 2002.

[9] http://spib.rice.edu/pib/select_noise.html, "Signal Processing Information Base (SPIB)",September, 2002.