6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

THE ANALYSIS AND APPLICATION OF A NEW ENDPOINT DETECTION METHOD BASED ON DISTANCE OF AUTOCORRELATED SIMILARITY[#]

Jie Zhu , *Fei-li Chen*

SJTU & Bell Labs Communications And Network Joint Laboratory
Shanghai Jiao Tong University , Shanghai, 200030, P.R.China
JieZhu@guomai.sh.cn , chenfeil@public1.sta.net.cn

Abstract

Endpoints detection play the important role in speech recognition. But the performance of some often used algorithms will decrease under low SNR conditions. A new concept as distance between autocorrelated functions is brought fully, and a new endpoint detecting method based on it is discussed in this paper. Results from experiment show that this new method has higher performance in endpoint detection. Especially in low SNR circumstance, it still can detect endpoint of speech accurately.

1. Introduction

Endpoint detection is very important in digital speech signal processing. Most methods used today are based on short-time energy. This method calculates the energy of noise and determines threshold. But when the energy of noise is unstable or SNR is low, the performance will decrease. A new endpoint detecting method based on distance between autocorrelated functions is discussed in this paper. This method has higher precision and it does not have special requests on noise model or energy, so it can detect endpoint accurately and efficiently even in low SNR circumstance.

2. Design Considerations

Speech signals are determined by frequency and amplitude but insensitive to phase. So we consider expressing speech signals by autocorrelated functions. If two random processes have the same autocorrelated function (or in some ratio), their power spectrums are the same (or similar) and the two speech signals are similar.

The auto-correlated function of a quasi-stationary random process is defined as below:

$$R(l) = \mathrm{E}[s(n)s(n+l)]$$

$$= \lim_{N\to\infty} \frac{1}{2N} \sum_{n=-N}^{n=N} s(n)s(n+l) \qquad (1)$$

If $s(n)$ is made up of a group of sine signals with frequencies ω_i , then

$$s(n) = \sum a_i \cos(\omega_i n + \phi_i) \qquad (2)$$

So we have $R(l) = \frac{1}{2}\sum a_i^2 \cos w_i l \qquad (3)$

In the case of end-point detection, speech signals are unstationary random processes, so we define short-time auto-correlated function as

$$R_w(l) = \frac{1}{N-l} \sum_{n=0}^{n=N-l-1} s(n)s(n+l) \qquad (4)$$

Its variance is defined as

$$\mathrm{D}[R_w(l)] = \mathrm{D}[\frac{1}{N-l} \sum_{n=0}^{N-l-1} s(n)s(n+l)]$$

$$= \frac{1}{(N-l)^2} \mathrm{D}[\sum_{n=0}^{N-l-1} s(n)s(n+l)]$$

$$= \frac{1}{(N-l)^2} \sum_{n=0}^{N-l-1} \mathrm{D}[s(n)s(n+l)] - \sum_{\substack{n,m=0;\\ n\neq m}}^{N-l-1} Cov[s(n)s(n+l), s(m)s(m+l)]$$

In most case the covariance between s(n)s(n+l) and s(m)s(m+l) can be ignored, so we have:

$$\mathrm{D}[R_w(l)] = \frac{1}{N-l} \mathrm{D}[s(n)s(n+l)] \qquad (5)$$

When $s(n)$ is white Gaussian noise with zero expectation and σ^2 variance,

$$\mathrm{D}[s(n)s(n+l)] = \sigma^4 \qquad (6)$$

So $\mathrm{D}[R_w(l)] = \frac{1}{N-l}\sigma^4 \qquad (7)$

When $l<<N$, the variance is small and become large when $l \to N$.

Assume two stationary random processes $s_0(n)$ and $s(n)$.Their short-time autocorrelated functions are $R_0(l)$ and $R_w(l)$ .Define

$$\lambda = \min_a \left[\frac{\sum_l [R_W(l) - aR_0(l)]^2}{\sum_l R_W{}^2(l)} \right] \qquad (8)$$

as distance between autocorrelated functions of these two random processes.

Let $\frac{\partial \lambda}{\partial a} = 0$, we have

$$a = \frac{\sum_l R_W(l) R_0(l)}{\sum_l R_0^2(l)} \qquad (9)$$

[#] This project was supported by Bell Labs(China), Shanghai.

Since λ represents the similarity distance between signals, we can use λ to detect endpoints. Assume the autocorrelated function of noise model is $R_0(l)$ and the input speech can be divided into two states: state 0 and state 1. State 0 represents the state without speech, which means $y(n)$=w(n), where w(n) means noise. State 1 represents the state with speech, which means $y(n)$=s(n)+$w(n)$, where s(n) means speech and w(n) means noise.

Assumed that the short-time autocorrelated function of y(n) is R(l), then in state 0, $R(l) = R_N(l)$,which $R_N(l)$ represents the short-time autocorrelated function of w(n).

According to (9), when in State 0, we have $a = \frac{\sum_l R_N(l)R_0(l)}{\sum_l R_0^2(l)}$, where $R_0(n)$ is the autocorrelated function of noise model. If w(n) is the same as (or similar to) the noise model, thus $\mathrm{E}[R_N(l)] = aR_0(l)$,λ reaches its minimum value with (9). So the signals being detected are most similar to noise model. According to the equation

$$\mathrm{D}[\mathrm{R}_\mathrm{N}(l)] = \mathrm{E}[\mathrm{R}_\mathrm{N}^2(l)] - (\mathrm{E}[\mathrm{R}_\mathrm{N}(l)])^2 \qquad (10)$$

so
$$\mathrm{E}[\lambda_0] \cong \frac{\sum_l \mathrm{E}[R_N(l) - aR_0(l)]^2}{\mathrm{E}[\sum_l R_N^2(l)]} = \frac{\sum_l \mathrm{D}[R_N(l)]}{\sum_l (a^2 R_0^2(l) + \mathrm{D}[R_N(l)])} \qquad (11)$$

From (7), with white noise, we have $\mathrm{D}[\mathrm{R}_N(l)] = \frac{a^2}{N-l}\sigma^4$, where σ^2 is the variance of noise model.

Because of $\sum_l R_0^2(l) = R_0^2(0) = \sigma^4$, if $l \in [0, \frac{N}{2} - 1]$• we have :

$$E[\lambda_0] = \frac{\sum_l \frac{1}{N-l}}{1 + \sum_l \frac{1}{N-l}} \qquad (12)$$

Its values $\in [1/3, 1/2]$. If the noise is not white, $\mathrm{E}[\lambda_0]$ is much smaller.

In State 1, $R(l) = R_s(l) + R_N(l)$. Where $R_s(l)$ means the short-time autocorrelated function of s(n). Here we divide $R_s(l)$ into two parts: $R_s'(l)$ and $\gamma \bullet R_0(l)$, where $R_s'(l)$ and $R_0(l)$ are orthogonal, and $\gamma = \frac{\sum_l R_s(l)R_0(l)}{\sum_l R_0^2(l)}$. So we have:

$$R_w(l) = R_s'(l) + R_N(l) + \gamma R_0(l)$$

As the same way in (9) and (11), there is the equation

$$\mathrm{E}[\lambda_1] \cong \frac{\sum_l \mathrm{R}_\mathrm{s}'^2(l) + \sum_l \mathrm{D}[R_N(l)]}{\sum_l [R_s'^2(l) + \mathrm{D}[R_N(l)] + (\gamma + a)^2 R_0^2(l)]} \qquad (13)$$

Use (7) in (13), we have

$$\mathrm{E}[\lambda_1] = \left(1 + \frac{\sum_l (\gamma + a)^2 R_0^2(l)}{\sum_l R_s'^2(l) + \sum_l \frac{a^2\sigma^4}{N-l}}\right)^{-1} \qquad (14)$$

In most case which the SNR is larger than 5dB and N>80 with l<N/2, we can estimate that, $\sum_l R_s'^2(l) >> \sum_l \frac{a^2\sigma^4}{N-l}$. so (14) is mainly determined by this part of

$$\phi = \frac{\sum_l R_0^2(l)(a+\gamma)^2}{\sum_l R_s'^2(l)} \qquad (15)$$

From $R_s(l) = R_s'(l) + \gamma R_0(l)$, we have :

$$\sum_l R_s^2(l) = \sum_l R_s'^2(l) + \sum_l \gamma^2 R_0^2(l) + 2\gamma \sum_l R_s'(l)R_0(l) .$$

Because of orthognonality relation we can know that the last term is zero, so we have:

$$\sum_l R_s'^2(l) = \sum_l R_s^2(l) - \sum_l \gamma^2 R_0^2(l) .$$

The equation (15) can be rewritten as

$$\phi = \frac{\sum_l R_N^2(l)}{\sum_l R_s^2(l)} \bullet \frac{(1 + \frac{\gamma}{a})^2}{1 - \frac{\gamma^2 \sum_l R_0^2(l)}{\sum_l R_s^2(l)}} \qquad (16)$$

From (16), we can find the first term $\frac{\sum_l R_N^2(l)}{\sum_l R_s^2(l)}$ is roughly proportional to the square of 1/SNR, and the second term is determined by the similarity of $R_s(l)$ and $R_0(l)$ .We define similarity ratio • as

$$\alpha = \frac{\gamma^2 \sum_l R_0^2(l)}{\sum_l R_s^2(l)} \qquad (17)$$

In fact, it represents the ratio of $R_s(l)$ to $R_0(l)$.

From above, we can find when SNR > 6db, the first term < 1/16, and when $\alpha = 0.3$•C $\frac{\gamma}{\alpha} = 1$, $\mathrm{E}[\lambda_1] > 0.8$.

With the analysis, we can conclude that when SNR is set, λ is mainly determined by the similarity of signals and noise. If they have the similarity ratio of 1, then $\mathrm{E}[\lambda_1] \to \mathrm{E}[\lambda_0]$. Because signals are consistent with noise model, we can not detect endpoints. Under common situations, the autocorrelated function of voiced sound takes on an obviously periodical distribution with large peak values. If noise does not have the same period, we have •<<1•similarity distance λ→1.

The autocorrelated function of unvoiced sound is similar to high frequency signals and does not have an obvious period. The similarity ratio • between unvoiced

sound and noise which has the similar spectrum structure is large, so we need higher SNR to reduce mistaken judgements. When SNR is high, the judgement will improve obviously. Generally, noise and unvoiced sound do not have the same frequence spectrum structure, we can get satisfactory result.

3. Realization of Algorithm And Determination of Distance Threshold

In practical application, we use 0.01 msec as the frame length for detection. We can know that the variance of a short-time autocorrelated function is determined by 1/(N-*l*).The larger *l* is, the larger variance is. We often use autocorrelated function of half frame as the length of auto-correlated function .

The realization algorithm is shown below:

1.Take a few of frames from the beginning of the input which have no speech, use equation (4) to find the autocorrelated function $\hat{R}(l)$, and make noise model $R_0(l)$.

2.Compute the short-time autocorrelated function of each frame.

3.Use (8) and (9) to calculate the distance between autocorrelated functions, so we get the parameter λ of each frame.

4.Estimate the threshold. We calculate the expectation E and variance σ of λ of the frames which is used to calculate the $\hat{R}(l)$, then decide the threshold.

5.Find out the endpoints of speech signals with the results of steps 3 and 4. E is usually used as beginning threshold. If • values of continuous a fixed number of frames are greater than the threshold, it means the beginning point. And E+σ can be used as the threshold of detection of ending. If the values of continuous 5 frames are less than it, it means the ending point.

In white noise condition, E is often between 0.3 and 0.4,and variance is about 0.2,which is consistent with analysis above.

Threshold is very important for endpoint detection. Different noise models have different threshold ranges. When SNR is high, threshold is kept stationary, otherwise threshold is determined by the statistical result of •.

4. Results Analysis

In experiments, we use method discussed above to detect the endpoints of more 300 utterances from males and females in different SNR environments and compare results with artificial methods. Some examples are analyzed as below:

(1) The detection results of single word with high SNR are shown in table 1.

From table 1,we can find autocorrelation method is quite good under higher SNR. Moreover, the mean error of autocorrelation method is 0.04 ms. In most case, the error is smaller than 0.03ms. But in detecting the speech beginning with unvoiced sounds, especially with the syllable [s], autocorrelated method is 3 to 4 frames slower at the beginning of the speech. The reason is that the signal power is very small at the beginning.

(2) The detection results of continuous speech with higher SNR are shown in table 2.

From table 2,it can be found that autocorrelation works well, especial in detecting the word gaps.

(3) The detection results of continuous speech with lower SNR are shown in table 3.

From table 3,we can find that the detection precision of autocorrelation method decreases when SNR is low, but the errors are less than 5 frames. Its performance is still satisfactory.

Through the experiments, we find autocorrelation method can accurately detect the endpoint of unvoiced sounds only when SNR is not less than 3db, otherwise the performance will decline a lot. In low SNR circumstance, errors often occur in syllables [n] and [ks]. When speech is continuos and middle syllables are weak, the results are affected more.

Tab.1 The examples of detection results of single word (SNR=15dB)

No	Auto-correlated	Artificial detection
1	202-247	195-249
2	200-235	198-236
3	167-213	163-215
4	165-195	165-196
5	299-322	299-322
6	198-233	190-234
7	187-237	184-240
8	216-268	210-268
9	208-282	202-267
10	192-228	192-230
11	190-251	190-252

Tab.2 The examples of detection results of continuous speech (SNR=20dB)

No	Auto-correlated	Artificial method
1	109-247, 264-394	109-247, 264-397
2	114-384	114-387
3	96-213,258-332, 338-372	106-215, 258-380
4	141-264, 291-448	141-264, 291-449
5	133-244,266-290, 300-391	133-244, 259-292, 300-390
6	124-255, 280-425	124-255, 276-425
7	145-270,321-420, 426-440	140-272, 299-441
8	115-252, 273-391	134-252, 271-392
9	117-254,299-356, 365-438	118-255, 299-440
10	160-267,303-351, 360-404, 413-432	160-270, 303-404, 413-433

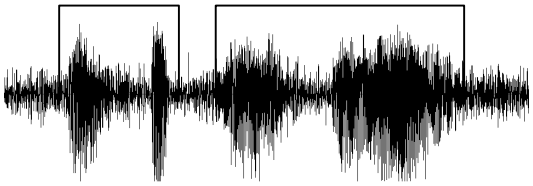
Tab.3 The examples of detection results of continuous speech (SNR<6dB)

No	Autocorrelated function method	Artificial detection
1	22-40,53-72,79-end	22-42,53-72,77-178
2	23-41,56-131	23-50,57-135
3	32-133,165-195 247-280	32-138,165-198 246-281
4	32-72	31-74
5	22-40,56-95	22-43,56-103
6	22-148	22-146
7	22-160,169-189	22-161,165-196
	201-253	198-260
8	30-70,101-148	30-75,96-149
9	38-55,69-149	37-60,65-155

5. Conclusions

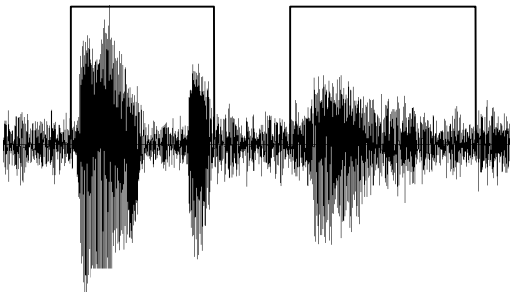
From above analysis, we can draw a conclusion that endpoint detection method based on distance between autocorrelated functions has higher performance, even when SNR is low, it still can detect endpoints of speech accurately.

Fig.1 The endpoint detection results of continuous speech with syllable [n] (SNR<6dB)



One six nine zero

Fig.2 The endpoint detection results of continuous speech with syllable [ks] (SNR<6dB)



four six two

6. References

1. Dermatas E.S. and Fakotakis N.D., etc, Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment. ICASSP, 1991, vol 4, pp733-736.

2. Savoji M.H., A Robust Algorithm for Accurate Endpointing of Speech Signals, Speech Commum. vol.8. pp45-60, 1989.

3. Rabiner L.R. and Faucon G., Study of a voice activity detector and its influence on a noise reduction system, Speech Commum, vol.16, pp 245-254,1995.

4. Doukas N. and Stathaki T. ,Naylor P., Speech Enhancement through nonlinear adaptive source separation methods. ICASSP 1996.

5. Doukas N. and Naylor P.,Voice Activity Detection Using Source Separation Techniques, Eurospeech, 1997.