International Symposium on Chinese Spoken Language Processing, pp. 327-330, Oct. 13-15, 2000, Beijing

(Student Excellent Paper Award)

A SELF-ADAPTING ENDPOINT DETECTION ALGORITHM FOR SPEECH RECOGNITION IN NOISY ENVIRONMENTS BASED ON 1/F PROCESS

WANG Fan, ZHENG Fang, WU Wenhu

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems,
Department of Computer Science & Technology, Tsinghua University, Beijing
fanwang@acm.org

ABSTRACT

This paper presents an effective and robust speech endpoint detection method based on 1/f process technique, which is suitable for robust continuous speech recognition system in variable noisy environments. The Gaussian 1/f process, which is a mathematical model for statistically self-similar random processes from fractals, is selected to model both speech and background noise. Then, an optimal Bayesian two-class classifier is developed to discriminate between real noisy speech and background noise by the wavelet coefficients with Karhunen-Loeve-type properties of the 1/f processes. Finally, for robust requirement, a few templates are built for speech and the parameters of the background noise can be dynamically adapted in runtime to deal with the variation of both speech and noise. In our experiments, 10 minutes long speech with different types of noises was tested using this new endpoint detector. A high performance with over 90% detection accuracy was achieved.

1. INTRODUCTION

Endpoint detection, which aims at distinguishing the speech and non-speech segments from digital speech signal, is considered as one of the key primary preprocessing components in automatic speech recognition (ASR) systems. Proper estimation of the start and end of the speech (vs silence or background noise) can avoid the wasting ASR evaluations on preceding or ensuing silence, which leads to efficient computation, and more importantly, to accurate recognition because misplaced endpoints cause poor alignment for template comparison. In some special but significant applications of ASR, the environments include many high level or nonstationary noises, such as in the mobile phone with speech command system. Noise comes from speakers (lip smacks, mouth clicks), environment (door slams, fans, machines) and transmission (channel noise, cross talk). The variability of durations and amplitudes for different sounds makes reliable speech detection difficult.

Some functions of signal short-time energy, zero-crossing rate or spectral energy have been conventionally used as the major features in the traditional endpoint detectors achieving good results for clean speech, but they fail considerably for speech with noises [1-3]. Pitch and entropy information are also chosen as the characteristics to distinguish speech signals from noisy signals, but the performances are dissatisfied, especially in the environments with high or nonstationary noises [4-5]. Other schemes use speech recognizer to determine the endpoints based on the output by a Viterbi algorithm by aligning the vocabulary word preceded and followed by a silence model or a noise model, but they require large computational resources [6].

The endpoint detection can be viewed as a speech/background-noise classification. For ASR systems, the ideal characteristics for such classification are: reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no a priori knowledge of the noise. Among these characteristics, robustness against noise conditions has been the most difficult to achieve. From the above directions of research, in this paper, the 1/f process, which is a mathematical model for statistically self-similar random processes from fractals, is chosen to model both speech and background noise, and a dynamically adapted endpoint detector by the wavelet coefficients of the 1/f processes is developed. The dynamics of airflow during speech and noise production may often result in some smaller or larger degree of turbulence. The geometry of these turbulence as reflected in the fragmentation of the time signal could be quantified by using 1/f process, and the wavelet basis expansions in terms of uncorrelated random variables are very good mathematic analyzing tools for such 1/f type behavior [7]. Therefore, in the new method, both the clean speech and background noise are modeled as two Gaussian 1/f processes. In addition, the wavelet coefficients of the real digital speech signal obtained in noisy environment are represented as the summation of the wavelet coefficients of these two 1/f processes. And then, an optimal Bayesian two-class classifier is developed to distinguish between the two Gaussian 1/f processes presenting real noisy speech and background noise by their robust wavelet coefficients with Karhunen-Loeve-type properties. Finally, for accurate and robust requirement, a few templates for speech are trained by clustering process, and the parameters of background noise can be dynamically adapted in runtime to deal with the variation of both speech and noise. In our experiments, 10 minutes long speech with different noises, whose average SNR is 10dB, was tested using the new endpoint detector. A high performance with over 90% detection accuracy was achieved.

2. 1/F WAVELET MODEL FOR SPEECH

The dynamics of airflow during speech production may often result in some smaller or larger degree of turbulence. Based on

the fractal theory, the signal of speech is a statistically self-similar random process whose statistics are invariant to dilations and compressions of the waveform in time [7]. More specifically, the speech signal $s(t)$ obeys the scaling relation with parameter H.

$$p(s(t)) = a^{-H} p(s(at)) \tag{1}$$

Where $p(s(t))$ denotes the probability density function of $s(t)$.

Based on equation (1), the mean and covariance functions of the statistically self-similar process are also self-similar. Such statistically self-similar process is generally defined as the *1/f process* having measured power spectra obeying a power law relationship of the form

$$S_s(\omega) \propto \frac{1}{|\omega|^{\gamma}}, \gamma = 2H + 1, 0 \le \gamma \le 2 \tag{2}$$

The wavelet basis expansions in terms of uncorrelated random variables constitute very good models for such 1/f-type behavior because the orthonormal wavelet basis expansions play the role of Karhunen-Loeve-type expansions for 1/f-type processes [8]. The 1/f process $s(t)$ can construct via such expansions

$$s(t) = \sum_m \sum_k s_k^{(m)} \varphi_k^{(m)}(t) \tag{3}$$

$$s_k^{(m)} =< s(t), \varphi_{mk}(t) >= 2^{\frac{m}{2}} \int s(t) \varphi(2^m t - k) dt$$

where the $\varphi_k^{(m)}(t)$ constitute a complete orthonormal set, $s_k^{(m)}$ are collection of mutually uncorrelated random variables which also called the wavelet coefficients of $s(t)$ with zero-means and variances via equation(2)

$$\begin{aligned} &E\left[s_k^{(m)}\right] = 0 \\ &\mathrm{var}[s_k^{(m)}] = \sigma^2 2^{-\gamma m} \end{aligned} \tag{4}$$

Consider the problem of endpoint detection of speech: the clean speech $s(t)$, background noise $n(t)$ and real speech with noise $\hat{s}(t)$ are separately modeled by Gaussian 1/f process via wavelet expansions

$$n(t) = \sum_{m \in M} \sum_{k \in N(m)} n_k^{(m)} \varphi_k^{(m)}(t)$$

$$s(t) = \sum_{m \in M} \sum_{k \in N(m)} s_k^{(m)} \varphi_k^{(m)}(t)$$

$$\hat{s}(t) = \sum_{m \in M} \sum_{k \in N(m)} \hat{s}_k^{(m)} \varphi_k^{(m)}(t)$$

$$t = 1, 2, \cdots T$$

where T is the number of sampling points. M represents the finite set of available distinct scales, and $N(m)$ is the set of available coefficients at each scale m. [8] shows that such wavelet-based representations are robust characterizations of 1/f-like behavior with Karhunen-Loeve-type properties.

In addition, exploiting the Karhunen-Loeve-type properties if the wavelet decomposition for 1/f-type processes, and using the fact that $s_k^{(m)}$ are independent of the $n_k^{(m)}$ and are decorrelated for any wavelet basis, the observation coefficients of real speech with background noise

$$\hat{s}_k^{(m)} = s_k^{(m)} + n_k^{(m)} \tag{5}$$

can be modeled as mutually independent zero-mean, Gaussian random variables with variance

$$\begin{aligned} &E\left[\hat{s}_k^{(m)}\right] = E\left[s_k^{(m)}\right] + E\left[n_k^{(m)}\right] = 0 \\ &\mathrm{var}\left[\hat{s}_k^{(m)}\right] = (\sigma_{\hat{s}}^m)^2 = \mathrm{var}\left[s_k^{(m)}\right] + \mathrm{var}\left[n_k^{(m)}\right] \\ &= (\sigma_s^m)^2 + (\sigma_n^m)^2 = \sigma_s^2 2^{-m\gamma_s} + \sigma_n^2 2^{-m\gamma_n} \end{aligned} \tag{6}$$

Then, for the speech/background-noise detection, an optimal Bayesian two-class classifier is built, which shows in Table 1.

Table 1: The description of the Bayesian two-class classifier of the speech/background-noise detection for the input signal $r(t)$.

Class	ID	Input Parameters	PDF	Parameters of PDF
Noise	H_0	$\left\{r_k^{(m)}\right\}$	Gaussian Probability Distribution	$\left\{(0, \sigma_n^m)\right\}$
Speech	H_1			$\left\{(0, \sigma_{\hat{s}}^m)\right\}$

Based on the Bayesian principle, the classifier can be represents as

$$\ln \frac{P(H_0 \mid \{r_k^{(m)}\})}{P(H_1 \mid \{r_k^{(m)}\})} = \ln \frac{P(\{r_k^{(m)}\} \mid H_0)}{P(\{r_k^{(m)}\} \mid H_1)} + \ln \frac{P(H_0)}{P(H_1)} \gtrless 0 \quad \begin{matrix} H_0 \\ H_1 \end{matrix} \tag{7}$$

where

$$P(\{r_k^{(m)}\} \mid H_0) = \prod_{m \in M} \prod_{k \in N(m)} p(\{r_k^{(m)}\} \mid H_0)$$

$$= \prod_{m \in M, k \in N(m)} \frac{1}{\sqrt{2\pi (\sigma_n^m)^2}} EXP\left(- \frac{\left(r_k^{(m)}\right)^2}{2\left(\sigma_n^m\right)^2} \right)$$

Hence,

$$\ln P(\{r_k^{(m)}\}|H_0) = -\frac{1}{2}\sum_{m\in M,k\in N(m)}\left\{\frac{(r_k^{(m)})^2}{(\sigma_n^m)^2}+\ln(2\pi(\sigma_n^m)^2)\right\} \quad (8)$$

Similarly,

$$\ln P(\{r_k^{(m)}\}|H_1) = -\frac{1}{2}\sum_{m\in M,k\in N(m)}\left\{\frac{(r_k^{(m)})^2}{(\sigma_{\hat{s}}^m)^2}+\ln(2\pi(\sigma_{\hat{s}}^m)^2)\right\} \quad (9$$

Equation (7) can be rewritten as via equations (8) and (9)

$$\ln\frac{P(H_0|\{r_k^{(m)}\})}{P(H_1|\{r_k^{(m)}\})} = -\frac{1}{2}\sum_{m\in N(m)}\left\{\left[\frac{1}{(\sigma_n^m)^2}-\frac{1}{(\sigma_{\hat{s}}^m)^2}\right](\sigma_r^m)^2+\ln\frac{(\sigma_n^m)^2}{(\sigma_{\hat{s}}^m)^2}\right\}+\ln\frac{P(H_0)}{P(H_1)} \quad (10)$$

where $(\sigma_r^m)^2 = \frac{1}{N(m)}\sum_{k\in N(m)}(r_k^{(m)})^2$ represents the variance of the input parameters, $(\sigma_n^m)^2$ 、 $(\sigma_{\hat{s}}^m)^2$ 、 $P(H_0)$、 $P(H_1)$ are prior knowledge.

3. ENDPOINT DETECTION ALGORITHM

3.1 Classifier Training

To obtain the parameter $(\sigma_{\hat{s}}^m)^2$ of real noisy speech, the parameter $(\sigma_s^m)^2$ should be trained before classification.

Because different type of speech has its own special characteristics in wavelet coefficients, a few templates for speech are trained separately to achieve high performance.

Assume $\{\{(\sigma_{sq}^m)^2 \mid m\in M\} \mid 1\le q\le Q\}$ represents the set of Q templates for clean speech

$$(\sigma_{sq}^m)^2 = \frac{1}{N(m)}\sum_{k\in N(m)}(s_k^{(m)})^2 = \frac{1}{N(m)}\sum_{k\in N(m)}<s_q(t),\varphi_k^m(t)>^2 (1\le q\le Q, 1\le t\le T_q) \quad (11)$$

where $s_q(t)$ means the training speech data with size T_q belonging to the template q $(1\le q\le Q)$.

A short period T' of background noise is first taken as the initial reference for the endpoint detection.

$$(\sigma_n^m)^2 = \frac{1}{N(m)}\sum_{k\in N(m)}(n_k^{(m)})^2 = \frac{1}{N(m)}\sum_{k\in N(m)}<n(t),\varphi_k^m(t)>^2 (1\le t\le T') \quad (12)$$

Another prior knowledge $\ln\frac{P(H_0)}{P(H_1)}$ is assumed to be equal to zero.

3.2 Adaptation in Runtime

In stable noisy environment, $(\sigma_n^m)^2$ is invariable. But to increase the detection accuracy in nonstationary noisy environments, the parameter of noise must be adapted in runtime.

The adaptation method is written as equation (13).

$$(\sigma_n^m)^2(t) = \frac{\sum_{i=1}^{t}(\sigma_n^m)^2(i)\cdot e^{-c\cdot i}}{\sum_{i=1}^{t}e^{-c\cdot i}} \quad (13)$$

Where t means the current speech segment is the t th background-noise segment, $(\sigma_n^m)^2(i)$ represents the variance of the wavelet coefficients at scale m for the i th noise segment. c a the exponential decay factor.

3.3 Algorithm

The input speech signal can be divided into F frames $\{R^i\} = \{\{r^i(1), r^i(2), \cdots, r^i(T_i)\}\}(1\le i\le F)$, and assume the first frame R^1 is the noise segment.

Step 1: Train the initial parameter of noise $\{(\sigma_n^m)^2\}$ via equation (12) using the data of R^1. Set the current frame number $i=1$.

Step 2: $i=i+1$. If $i>F$, the algorithm ends, unless compute the set of wavelet coefficients $\{r_k^{(m)} \mid m\in M, k\in N(m)\}$ of frame R^i.

Step 3: Compute the variance of the wavelet coefficients at each scale.

$$(\sigma_r^m)^2 = \frac{1}{N(m)}\sum_{k\in N(m)}(r_k^{(m)})^2$$

Step4: For every template of clean speech, use equation (10) to classify with $\{(\sigma_r^m)^2\}$, $\{(\sigma_{sq}^m)^2\}$ and $\{(\sigma_n^m)^2\}$. If R^i is

classified as speech segment with *any one* template, R^i is just labeled as speech segment, then go to Step 2; unless, R^i is labeled as background noise segment.

Step 5: The parameter of noise $\left\{(\sigma_n^m)^2\right\}$ is update via equation (13). Go to Step 2.

When all frames are labeled as speech or background-noise segments, the endpoint boundaries can be found in the change points of segment type. Finally, some boundary pairs with the period of the corresponding speech segment less than a predefined minimum duration are rejected.

4. EXPERIMENT RESULTS

The speech database used in the experiments is a 20 minutes speech stream with about 300 Mandarin Chinese sentences from 10 speakers (native Chinese males and females) in clean environment. The sampling rate is 16KHz with 16bits resolution. The first half of the database is used as training data to get the parameters of clean speech, and the last half one is added randomly with some different types of noises and tested by the new algorithm to obtain its performance. The noises in the test corpus are from speakers and environments, such as white noise, pink noise, breath, mouth clicks and door slams, but don't include the background speech noise such as another speaker's speech or transmission noise. The level of noise is variational with average SNR of 10dB. In processing, the frame size is 16ms with floating size of 8ms. The standard speech/background-noise classifications are labeled by manual. In the experiments, three clustering methods are tested to train the parameters of clean speech:

1. Clean speech is classified as 4 templates, including
   - Unvoiced consonant
   - Voiced consonant
   - Vowel
   - Transient region between consonant and vowel

   Label the training data by manual into above four types, and train the parameters separately.
2. Clean speech is classified as 4 templates, and labeled the training data by Vector quantization (VQ) method.
3. Clean speech is classified as 10 templates, and labeled the training data by Vector quantization (VQ) method.

The experiments results in showed in Table 2.

Table 2: The classification accurate rate of the test data by the new speech/background-noise detection method based on the 1/f process.

Training Method	4 templates labeled by manual	4 templates labeled by VQ	10 templates labeled by VQ
No Adaptation	85.6%	78.2%	80.7%
Adaptation in Runtime	92.0%	89.8%	90.5%

It can be noted from Table 2 that the new endpoint detection algorithm obtains high performance to deal with nonstationary noisy environments, which is suitable for ASR systems in noisy conditions. The self-adaptation in runtime for the parameters of background noise is very useful for nonstationary noisy conditions, which increase over 10% accurate rate. Other fact is that increasing the number of templates for clean speech could not result in high increase of accurate rate because more templates will cause more errors in clustering.

5. CONCLUSION

A 1/f process based self-adapting endpoint detection algorithm is presented in this paper. It's worthwhile to point out three advantages of the proposed method in comparison with other existing algorithms for speech endpoint detection. First, because of the production principles of speech and noise, the 1/f process is chosen to accurately and compactly model the speech and background noise and the robust wavelet coefficients with Karhunen-Loeve-type properties of the 1/f processes are used as the features for classifier to detection, which can represent the characteristics in different multi-resolutions of speech and noise. Second, the multi-templates for clean speech and dynamical parameters adaptation for background noise in runtime will keep high performance in both variable background noises and speak styles. Third, none threshold and prior knowledge of noise is required in the new method. Future work will combine the speech endpoint detection together with the denoise methods based on 1/f process theory.

6. REFERENCES

[1] Lamel L., Labiner L., Rosenberg A. and Wilpon J. "An Improved Endpoint Detector for Isolated Word Recognition". *IEEE ASSP Magazine*, 29:777-785, 1981.

[2] Savoji M.H. "A Robust Algorithm for Accurate Endpointing of Speech". *Speech Communication*, 8:45-60, 1989.

[3] Junqua J.C., Mak B. and Reaves B., "A Robust Algorithm for Word Boundary Detection in the Presence of Noise". *IEEE Trans. on Speech and Audio Processing*, 2(3):406-412, 1994.

[4] Hanada M., Takizawa Y. and Norimatsu T. "A noise robust speech recognition system". *Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing*, 893-896, 1990.

[5] Shen J.L., Hung J.W. and Lee L.S. "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments". *International Conference on Spoken Language Processing*, Sydney, Nov. 1998.

[6] Wilpon J.G. and Rabiner L.R. "Application of hidden Markov models to automatic speech endpoint detection". *Comput. Speech Language*, 2:321-341, 1987.

[7] Manderbrot, B.B. *The Fractal Geometry of Nature*, Freeman, 1982.

[8] Wornell, G. "Wavelet-based representation for the 1/f family of fractal process." *Proceeding of IEEE*, 81:1428-1450, 1993.