A FRACTAL BASED VOICE ACTIVITY DETECTOR FOR INTERNET TELEPHONE

Su Yang, Zong-Ge Li, Yan-Qiu Chen

Department of Computer Science and Engineering & Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China
E-mail: suyang@fudan.edu.cn

ABSTRACT

The fractal dimension is useful in discriminating different textures of images. Also, the distinction between the 'texture' of speech signals and that of noises should exist because speech signals are regular 'man-made' processes and noises are random 'natural' processes. Through experiments, we found that the fractal dimension of speech signals is distinctly different from that of noises. Thus, a fractal based voice activity detector (VAD) is proposed. The experimental results show that the proposed VAD outperformed two traditional approaches based on energy and zero-crossing rate at low SNR. Applied to the Internet telephone developed by us, this VAD improves the quality of service.

1. INTRODUCTION

As communication costs have dropped greatly by using Internet telephony, increasingly many people prefer Internet telephony for long distance calls. As Internet telephony is based on IP networks, speech signals must be transformed to data packets before being transmitted to the remote end, where speech signals are restored from received data packets. Because the original goal to develop IP networks is to provide services for data communication, for real-time speech communication such as Internet telephony, the quality of service (QoS) cannot be fully guaranteed. To improve the performance, voice activity detection (VAD) is an essential method [1][2]. As the Internet telephone works in duplex mode, the local user usually keeps silent while listening to the remote user. At that time, ambient noises at the local end are captured and transmitted to the remote end through Internet. In this case, network resources are consumed by meaningless noises. For silence often takes place and covers 60% duration in a conversation, we need the so-called voice activity detector, by which unvoiced frames can be removed and only voiced frames are transmitted to the remote end. Traditional VAD methods are based on energy and zero-crossing rate [1][2]. As their performance is not satisfactory at low SNR [3], investigations for new approaches are ongoing so far. [4]. Through experiments, we found that the fractal dimension of voiced frames is distinctly different from that of unvoiced frames. The fractal dimension is useful in describing the texture of images. From a similar viewpoint, it is also useful in describing the 'texture' of signals. As speech signals can be regarded as 'man-made' processes and noises can be deemed as 'natural' processes, the texture of speech signals should be different from that of noises. This motivates us to utilize the fractal dimension for voice activity detection. Here, we employ the blanket-covering dimension among the many definitions of fractal dimension because it has some properties fitting well into the general requirements for feature extraction [5][6]. Then, we applied the proposed VAD to a PC to PC Internet telephone developed by us. Experimental results show that this VAD improves the QoS of the Internet telephone in comparison with the case without VAD. The evaluations also show that the proposed VAD achieved better performance at low SNR than the two traditional methods [1][2].

2. VOICE ACTIVITY DETECTION BASED ON FRACTAL DIMENSION

2.1. Speech Analysis with Fractals

As a geometrical tool, the fractal dimension can characterize waveform details of a signal. Among the many definitions of fractal dimension, the blanket-covering dimension is very suitable for feature extraction [5], because it is invariant to signal shifting in the coordinates of time and amplitude [6]. For a time sequence $f(n)$:n=0,1,…,N, choosing a scale r to form an upper envelop $U_r(n)$ and a lower envelop $L_r(n)$, where

$$U_r(n)=\max\{U_{r-1}(n-1), U_{r-1}(n)+1, U_{r-1}(n+1)\}, \quad (1)$$

$$L_r(n)=\min\{L_{r-1}(n-1), L_{r-1}(n)-1, L_{r-1}(n+1)\}, \quad (2)$$

$$U_0(n)=L_0(n)=f(n), \quad (3)$$

the fractal measurement can then be calculated by

$$L(r)=\frac{1}{2r}\sum_{n=0}^{N}[U_r(n)-L_r(n)]\,. \qquad (4)$$

The fractal dimension at a given scale r is

$$D_r=1-[\log L(r+1)-\log L(r)]/[\log(r+1)-\log r]. \qquad (5)$$

Figure 1 shows the fractal dimension at scale 36 of some voiced and unvoiced frames that were acquired at 8kHz sampling rate and 16-bit resolution. The duration of each frame is 0.5 second. The choice of 0.5 second will be explained later. In Fig. 1, the fractal dimensions of the voiced frames are at the top while those of the unvoiced frames are at the bottom. This indicates that the fractal dimension of voiced frames is greater than that of unvoiced frames. The distinction between the two classes is obvious. Motivated by the above experimental results, the decision on whether a frame is voiced or not is made by comparing its fractal dimension with a given threshold. For this, we should obtain three parameters, T, R, and F, at first. Here, T is the threshold with which the fractal dimension of each frame is compared, R is the scale at which the fractal dimension is used for comparison, and F=1 or -1 indicates whether the fractal dimension of voiced frames is greater than that of unvoiced frames or not, respectively. Then, we hold:

$$\begin{array}{l}\text{When } F=1,\text{ if } D_R>T,\text{ voiced.}\\ \text{When } F=-1,\text{ if } D_R<T,\text{ voiced.}\end{array} \qquad (6)$$

Although the three parameters may vary with microphones, sound cards, and computers, we found that the optimal scale is around 30 for most computer configurations. Note that we need such preprocessing before computing the fractal dimension that each frame should be normalized to possess unit energy by dividing the square root of its energy.

2.2. Performance Comparisons

Misclassification rate of the proposed method was evaluated at different SNR and compared with those of two traditional methods based on energy and zero-crossing rate [1][2]. The detailed procedure is as follows:

(1) It is assumed that there are N voiced and unvoiced frames denoted as $[V_1,V_2,\ldots,V_N]$ and $[U_1,U_2,\ldots,U_N]$ respectively. Let $S_i=V_i+t\times U_i$ and $X_i=t\times U_i$, where $i\in[1,N]$ and t is the factor to fulfill a given SNR. The SNR is defined as $10\mathrm{Log}_{10}\{\sum E(V_i)/\sum E(X_i)\}$, where $E(V_i)$ and $E(X_i)$ represent the energy of V_i and X_i respectively. Let $F(S_i)$ and $F(X_i)$ represent the feature value of S_i and X_i respectively. Corresponding with which VAD method is used, the feature value of S_i and X_i is one of the following three, the blanket-covering dimension at a given scale, energy, and zero-crossing rate. By sorting both $\{F(S_i)|i\in[1,N]\}$ and $\{F(X_i)|i\in[1,N]\}$ in ascending order, we can obtain two sequences denoted as $[F'(S_1),F'(S_2),\ldots,F'(S_N)]$ and $[F'(X_1),F'(X_2),\ldots,F'(X_N)]$ respectively. Let $F'(S_M)$ and $F'(X_M)$ be two cluster centers, where $M=N/2$.

(2) If $|F(S_i)-F'(S_M)|<|F(S_i)-F'(X_M)|$, we classify S_i into the voiced group and label it with $G(S_i)=1$. Otherwise, S_i is classified into the unvoiced group and labeled with $G(S_i)=0$. The criteria to classify X_i into either the voiced or the unvoiced group is similar. By this classification, we can obtain two groups.

(3) Let $P=N-\sum G(S_i)$ and $Q=\sum G(X_i)$. P is the number of the voiced frames that are classified into the unvoiced group. Q is the number of the unvoiced frames that are classified into the voiced group. The overall misclassification rate is

$$W=(P+Q)/2N. \qquad (7)$$

Table 1 shows the misclassification rate of the three methods at different SNR, where Fractal denotes the proposed method, Energy the traditional method based on energy, and Zero the traditional method based on zero-crossing rate. Obviously, the performance of the proposed method is better than the other two.

Table 1: Misclassification rate (%) versus SNR (dB)

SNR	FRACTAL	ENERGY	ZERO
5	6.82	14.2	15.34
4	7.39	14.2	15.91
3	7.95	14.2	16.48
2	9.09	14.77	17.61
1	10.23	15.34	20.45
0	11.36	16.48	21.59

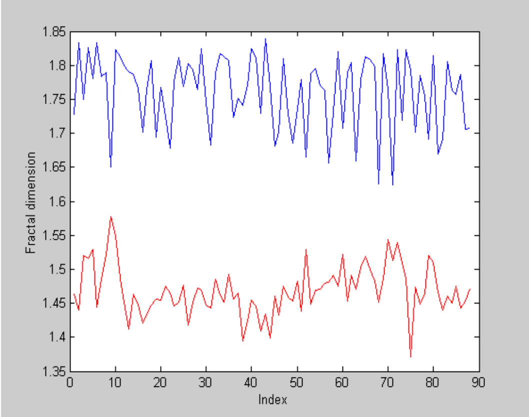

Fig. 1: Fractal dimension of voiced and unvoiced frames

3. APPLICATIONS

3.1. Threshold Determination

The fractal based voice activity detector has been applied to Internet telephony to improve its QoS. Before the voice activity detector works, the previously mentioned three parameters, T, R, and F, should be determined by the following procedure.

(1) Let the user read an article and then keep quiet. Capture both the voiced and unvoiced part into computer. Divide the voiced and unvoiced part into N frames respectively, which are denoted as $\{S_1,S_2,\ldots,S_N\}$ and $\{U_1,U_2,\ldots,U_N\}$. Here, the length of each frame is equivalent.

(2) Let r=1, where r is the scale to compute the fractal dimension.

(3) Supposing that $D_r(S_i)$ and $D_r(U_i)$ are respectively the fractal dimension of S_i and U_i at scale r, compute $\{D_r(S_1), D_r(S_2) ,\ldots, D_r(S_N)\}$ and $\{D_r(U_1), D_r(U_2),\ldots, D_r(U_N)\}$.

(4) Produce the histogram of set $D=\{D_r(S_1),D_r(S_2),\ldots, D_r(S_N);\ D_r(U_1),D_r(U_2),\ldots,D_r(U_N)\}$. Note that the histogram is produced on the basis of both the voiced and the unvoiced frames. Actually, fractal dimension must be within [1,2] in two-dimensional space. Divide [1,2] into K bins and denote the ith bin as B_i, where B_i is $[1+(i\text{-}1)/K,1+i/K]$. Let N_i represent the number of such elements in set D that lie in B_i. Then, N_i versus B_i forms the histogram, i=1,2,…,K. If two dominant peaks as shown in Fig. 2 exist, then, it should be able to distinguish the voiced frames from the unvoiced frames. If the fractal dimension of the voiced frames is greater than that of the unvoiced frames, let F_r=1. Else, let F_r=-1. If we cannot find two dominant peaks at all, then, let F_r=0 and go to (7).

(5) Supposing that two dominant peaks lie in B_i and B_{i+M}, the index of the bottom between the two dominant peaks is I=arg min$\{N_i,N_{i+1},\ldots,N_{i+M}\}$. Then, the threshold T_r is defined as the left or right boundary of B_I. If F_r=1, T_r=1+(I-1)/K. If F_r=-1, T_r=1+I/K.

(6) Compute the misclassification rate W_r as defined in Eq. (7).

(7) If r<=$R^{'}$, where $R^{'}$ is a predefined value, r=r+1; go to (3). Else, go to (8).

(8) Finally, let the three parameters be R=arg min$\{W_r|r=1,2,\ldots,R^{'}\}$, T=T_R, and F=F_R.

3.2. Speech Communication on Internet

We developed a PC to PC Internet telephone using VC++. This software contains the following modules: sound acquisition, playback, transmission, encoder, decoder, and voice activity detector. The architecture of this Internet telephone is illustrated in Fig. 3. The data transmission is based on TCP/IP. Experiments were conducted on public networks, where two users were connected to Internet from Beijing and Xi'an respectively. The distance between the two cities is above 1000 kilometers. In the experiments, it can be perceived that the QoS was improved by running the proposed VAD in comparison with the case without VAD.

To optimize the overall performance of the software, we let the duration of each frame be 0.5 second. As this Internet telephone is based on TCP/IP, an IP head and a TCP head must be added to each data packet carrying a frame. If the duration of each frame is too short, the number of total data packets will increase. Correspondingly, the QoS will decrease because more heads will be added. On the contrary, as the length of each frame becomes longer, the delay time will increase. Besides, it may become difficult to understand a sentence if one frame is lost. So, a compromise is needed. According to our experience achieved from experiments, we let the length of each frame be 0.5 second.

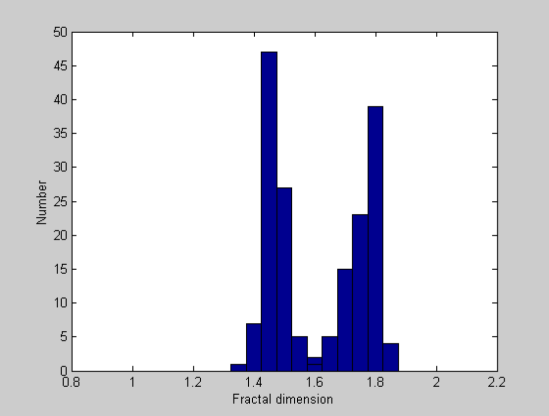

Fig. 2: Histogram with respect to fractal dimension of voiced and unvoiced frames

4. SUMMARY

In this paper, we proposed a fractal dimension based voice activity detector because we found that the fractal dimension of speech signals is distinctly different from that of noises in correspondence with silence. At low SNR, the performance of the proposed method is better than that of the method based on energy and zero-crossing rate. The proposed method has been applied to a PC to PC Internet telephone developed by us. The QoS is improved by integrating the proposed VAD. This proves that the proposed VAD is effective in silence compression.

REFERENCES

[1] L. R. RABINAR and M. R. SAMBUR, "An algorithm for determining endpoints of isolated utterances", Bell Syst. Tech. J., Vol. 54, pp. 297-300, 1975.

[2] Benyassine etc., "ITU-T recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications", IEEE Communication Magazine, pp. 64-73, September 1997.

[3] S. Hatamian, "Enhanced speech activity detection for mobile telephone", Proc. IEEE 42nd Veh. Technol. Conf., pp. 159-162, 1992.

[4] C.-H. Liu and Huang C.-C., "Voice activity detector based on CAPDM architecture", Electronics Letters, Vol. 37, No. 1, pp. 68-69, 2001.

[5] S. YANG, Z.-S. LI, and X.-L. WANG, "Ship recognition via its radiated sound: The fractal based approaches", Journal of the Acoustical Society of America, Vol. 112, Issue 1, pp. 172-177, 2002.

[6] P. MARAGOS and F.-K. SUN, "Measuring the fractal dimension of signals: Morphological covers and iterative optimization", IEEE Transactions on Signal Processing, Vol. 41, pp. 108-121, 1993.

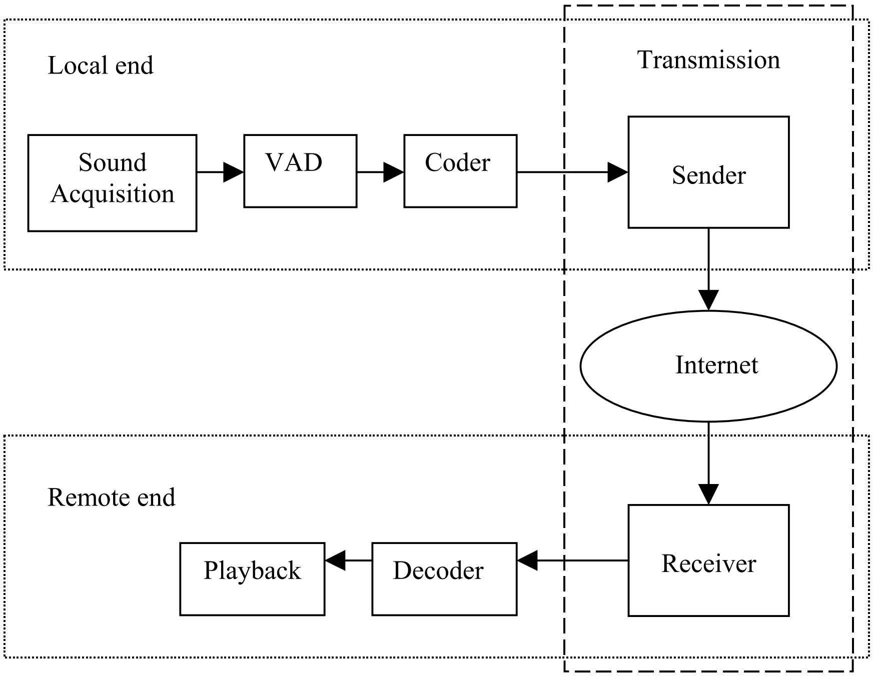

Fig.3: Architecture of the PC to PC Internet telephone