A Silence/Voice Segment Detection Method of Speech Signal Using Wavelet Transform Parameters

Jianping Xie，Jiandong Zhou
School of Computer and Information Engineering, Lishui University, Lishui 323000, Zhejiang P. R. China
E-mail:xjp1386@hotmail.com,zjd950131@sina.com

Abstract

As to the problem of low accuracy of the end detection in the commonly-used detection method of the speech end in recent years, this paper proposes a silence/voice detection method of speech signal using wavelet transform parameter. Using wavelet's ability of frequency segmentation and energy focusing, the statistic parameters of the speech signals on different subbands are extracted. Then importing parameters validity analysis based on fuzzy entropy, we get the most discriminable and stable parameters of different subbands as the discrimination parameters. Simulation experiment results prove this method is stable and effective under different noise conditions, and get some improvements in precision and robustness compared with the method based on the traditional parameters.

Key Words: Speech signal; Wavelet transform; Data compression; Fuzzy entropy; Simulation experiment

1. Introduction

The speech signals are the important information carrier, and the speech communication is the important constituting parts of the modern communication. According to the statistic results, about half or more of the time is in the silent condition during man's common speech [1]. It's necessary to retain the silent time slots for the stored medium like the CD level. But it's a waste for the speech communication to retain so much silent time slots. In the traditional speech communication, the coding method is utilized to eliminate this kind of information redundancies by means of the correlations between these silent time slots. However, it's not enough to use this method to restrain the silent time slots. For this problem, the detection method of speech end is usually used to locate the start and end of the speech and then eliminate the silent time slots in recent years [3, 4]. Nevertheless, the validity of the detection parameters used in this method is limited and so leads to a low accuracy in the end detection, its effect is even much worse especially when the signal-to-noise ratio is lower.

In order to solve the above problems, the detection method of the silent speech slots and the speech slot based on wavelet transform parameters is proposed in this paper. First the wavelet transform is carried out for the speech signals, the statistical feature of the wavelet coefficient is calculated after obtaining the wavelet transform coefficient of the signals. The statistical parameters of the wavelet coefficient in different frequencies levels is considered to be the differentiating parameters of this frame of signal, and whether it belongs to the silent time slot or the speech slot is then determined. The wavelet transform possesses high performance of frequency segmentation and energy focusing, thus the signals features of the silent time slot and the speech slot in different frequencies bands and time slots are "projected" in the wavelet coefficient of different levels, and the parameters like these have stronger performance in discriminating the silent slot and the speech slot. Therefore, the accuracy and robustness of this method are increased to a large extent compared with the traditional parameters.

2. The wavelet transform theory

The continuous wavelet transform is as follows

$$CWT(a,b) = \int_{-\infty}^{+\infty} f(t)\overline{\psi_{a,b}(t)}dt \quad (1)$$

$$\psi_{a,b}(t) = |a|^{-1/2}\psi(\frac{t-b}{a}) \quad (2)$$

where $\psi(t)$ is the wavelet function, a is the scale factor and stands for the frequency graduation after the signals are transformed, b is the translating factor and stands for the time graduation after the signals are transformed. $\psi(t)$ is a locally supporting function, so the wavelet transform has such performance as local analysis compared with the Fourier transform, and it's suitable for non-stationary signals. In the mean time, $\psi_{a,b}(t)$ has such performance as time-frequency analysis changing with scale factor. The frequency resolution of time-frequency unit increases with the increment in the scale factor, which denotes the decrement in the frequency element; the frequency resolution of time-frequency unit decreases with the decrement in the scale factor, which denotes the increment in the frequency element, this is the strongest superiority

This project was supported by the new century reform project in education and teaching of Zhejiang Province (yb09072)

978-0-7695-4413-7/11 $26.00 © 2011 IEEE
DOI 10.1109/DMDCM.2011.72

compared with the short-time Fourier transform. If the above wavelet transform satisfies the condition $C_\psi = \int_R \frac{|\psi(w)|^2}{|w|} dw < \infty$, there exists the following inverse transform:

$$f(t) = C_\psi^{-1} \int_0^\infty \int_{-\infty}^{+\infty} CWT(a,b) \psi_{a,b}(t) \frac{dadb}{a^2} \quad (3)$$

In the engineering conditions, the existing condition is approximately equivalent to $\int_{-\infty}^{+\infty} \psi(t)dt = 0$, and it is easily satisfied. From the above formula, it can be seen that both the normal transform and the inverse transform can be implemented for the wavelet function if it satisfies the admitted condition. In this paper, the "normality" condition is added to $\psi(t)$ so as to make the wavelet function possess stronger localization property in frequency domain and increase its functions of frequency grade and energy focus. So we put forward a "normality" condition for $\psi(t)$, that is

$$\int t^p \psi(t) dt = 0, p = 1 \sim n \quad (4)$$

the bigger n is, the better it is.

This condition is equivalent to that $\psi(t)$ has high-order zero at $\omega = 0$ in the frequency domain, and the higher the order is, the better it is (The first-order zero is the admitted condition). It can be proved that $CWT(a,b)$ will decrease with at the speed of no less than $a^{n+3/2}$ with this condition. Therefore, the wavelet function has higher localization property in the frequency domain. In the mean time, this condition diminishes the contribution of the items of the expanded formula whose orders are less n in the corresponding time domain, the high-order variation of the signals becomes more outstanding and the "energy focus" function of the wavelet transform is increased.

The continuous wavelet transform is discretized by means of Mallat algorithm. Mallat algorithm is proposed according to multi-resolution analysis, it's essentially a series of subspace in L2(R) satisfying a certain conditions. That is, the space L2(R) is divided into a series of subspace $\{V_j, j \in Z\}$ by the multi-resolution analysis, and it satisfies the condition $V_j \subset V_{j-1}$. The orthocomplementation of V_j is W_j, and it satisfies $V_{j-1} = W_j \oplus V_j$. On the basis of dividing the space like this, Mallat proves that a group of orthogonal basis in W_j can be constructed and it satisfies the requirements of the wavelet function. According to the above multi-resolution analysis theory, the signals is projected to a series of subspace of V_j and W_j, where V_j constitutes the low-pass area and W_j constitutes the band-pass area. According to the multi-resolution analysis theory, there exists $A_{j-1} f(x) = A_j f(x) + D_j f(x)$, where $A_j f(x)$ and $D_j f(x)$ are the projections of the signals in V_j and W_j, respectively. The basic decomposition algorithm is

$$\begin{aligned} a_n^j &= \sum_{k=-\infty}^{+\infty} h_{k-2n} a_k^{j-1} \\ d_n^j &= \sum_{k=-\infty}^{+\infty} g_{k-2n} a_k^{j-1} \end{aligned} \quad (5)$$

The reconstruction formula is

$$a_n^{j-1} = \sum_{k=-\infty}^{+\infty} h_{n-2k} a_k^j + \sum_{k=-\infty}^{+\infty} g_{n-2k} d_k^j \quad (6)$$

3. The discrimination method of the silent segment and the speech segment

From the above discussion, it can be seen that wavelet transform is suitable for nonstationary signals analysis, and this is in agreement with the property of the speech signals. In addition, the wavelet transform has such function of "energy focus". Therefore, after the wavelet transform is complemented for the speech signals, some features of the speech segment will be enhanced in different "sub-space" of wavelet transform, and characteristics of the speech segments in different "sub-space" will get more detailed characterization. The method used in calculating the statistical properties of signal transform coefficients in the sub-space will be more accurate in differentiating silence/speech segments compared with the method barely calculating the statistical properties of the time domain and frequency-domain. Accordingly, we propose a method detecting the silent segments and the speech segments based on wavelet transform. The flowchart of the algorithm is shown in Figure 1.

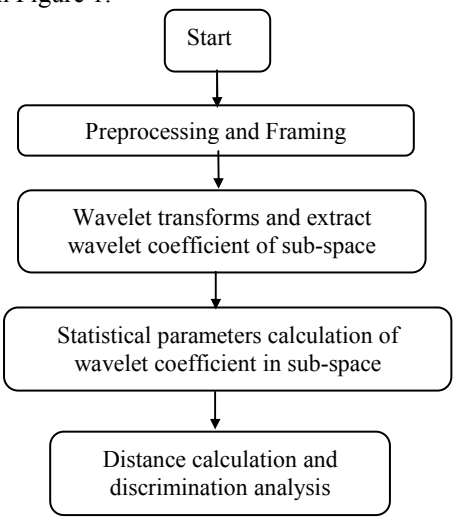

Figure 1. Flowchart of the detection algorithm

Preprocessing includes signal filtering and other process for regulating, and consistent with the general process of speech processing. After processing, the speech signals need to be framed, generally about 33~100 frames per second.

In order to maintain a smooth transition, there is general overlap between frames, frame shift and frame size ratio generally is selected to be 0 to 0.5. Signals framing is

used to ensure a signal in a stationary. The length of the frame is very important for the traditional method of speech signal processing. Wavelet transform was originally non-stationary signal analysis tool, so the choice of frame size has little effect. Shubha Kadambe and G..Faye Boudreaux-Bartels have discussed this in detail [5].

The wavelet coefficient of the subspace is obtained when the speech signals go through wavelet transform frame-by-frame after framing. Considering the calculation cost, Mallat algorithm is utilized to complete wavelet decomposition of the three-tier using db3 wavelet. Thus an approximate coefficient ca3 and three detail coefficients cd3, cd2, and cd1 in the subspace are obtained. It's necessary to thoroughly analyze them in order to compare their performances in discriminating the silent segments and the speech segments of 12 parameters in the wavelet domain.

The main statistical parameters of the speech signals include amplitude, energy, quasi-periodicity, and zero-crossing rate, etc. The traditional method of discriminating silent segments and the speech segments depends on these statistical features. The selection of the statistical parameters in the wavelet subspace borrows ideas from the traditional analysis method in the time domain. The traditional statistical parameters mainly include mean value M, variance B, average zero passage rate Z, which are defined to be

$$M = \frac{1}{N}\sum_{i=1}^{N} x_i \quad M = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{7}$$

$$B = \frac{1}{N}\sum_{i=1}^{N} (x_i - M)^2 \tag{8}$$

$$Z = \frac{1}{2N}\sum_{i=1}^{N} \left|\mathrm{sgn}[x_i] - \mathrm{sgn}[x_{i-1}]\right| \tag{9}$$

where $x_i\ (i = 1,2,\cdots,N)$ is the sample data, N is the sample number in a frame. Considering the different features of the subspace coefficient after the signals are wavelet transformed, the statistical data which are effective in the time domain are not always effective in the wavelet subspace, and it's same in the low-frequency and the high-frequency subspace. So the analysis method of parameters validity based on fuzzy entropy is introduced to thoroughly analyze the discrimination performance of the silent/speech segments of the 12 parameters obtained from the wavelet domain. The silent and speech segments can be regarded as two fuzzy sets for their uncertainty. The membership function are defined for the discrimination parameters on the two fuzzy sets, the membership function degree of every parameter in the fuzzy sets are used to measure the membership degree of a frame of signals in the set. Then the information entropy of the parameter membership is calculated. The bigger the entropy is, the higher the uncertainty of the parameter in the fuzzy sets, the lower the discrimination performance is, and vice verso. Next is the concrete method.

First define two discourse domains P={Parameters belong to the speech segment} and Q={Parameters belong to the silent segment}, then define the membership function for the parameter to follow Cauchy distribution

$$\mu(x) = \frac{1}{1 + \alpha(t - m)^{\beta}} \tag{10}$$

where m is the mean value of t, β=2, α is the reciprocal of variance.

The above formula reflects the membership grade of the parameters in two discourse domains, and the smaller the characteristic parameters in the discourse domain are, the higher the membership grade in the discourse domain is. The fuzzy entropy is defined in the above fuzzy sets

$$H_{j,k} = S(\mu_{j,k}(t)) = -\mu_{j,k}(t)\ln(\mu_{j,k}(t)) - (1-\mu_{j,k}(t))\ln(1-\mu_{j,k}(t)) \tag{11}$$

where j=1, 2, …,12; k=1,2. j denotes the j^{th} of the 12 parameters, k denotes the silent and speech segments. According to formula (11), the fuzzy entropy of the j^{th} parameter in the k^{th} fuzzy set can be gained. For N samples, their average value is used to be the entropy evaluation of the j^{th} parameter in the k^{th} fuzzy set.

$$\overline{H}_{j,k} = \frac{1}{N}\sum_{i} H_{j,k}(t^i) \tag{12}$$

Synthesize the entropy of the j^{th} parameter in two fuzzy sets, the effective measuring of the j^{th} parameter in the fuzzy set can be obtained after normalization. The bigger V is, the higher discrimination capacity of the j^{th} parameter in the fuzzy set, and the more effective it is.

$$V_j = \frac{\sum_{k=1}^{2} 1/\overline{H}_{j,k}}{\sum_{j=1}^{12}\sum_{k=1}^{2} 1/\overline{H}_{j,k}} \tag{13}$$

In accordance with the above scheme, three groups of signals are selected and 100 samples are obtained, and the Gaussian white noises are added to them. The wavelet statistical parameters of the signals are calculated with the signal-to-noise ratio 20dB, 15dB, 10dB, 5dB, 0dB, respectively, and then the validity is analyzed. The results are shown in Figure 2.

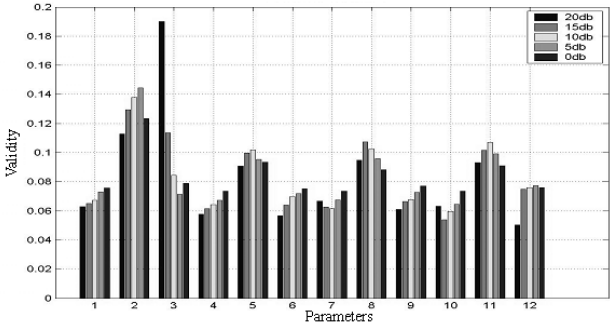

Figure 2. The discrimination capability of the wavelet parameters in silent/speech segment

12 parameters are the mean value, variance, and zero-crossing rate of ca3, cd3, cd2, and cd1. From Figure 2,

it can be seen that the validity of the 2nd, 3rd, 5th, 8th, and 11th parameter are obviously higher than that of other parameters. The discrimination capability of zero-crossing rate of ca3 is obviously higher than that of other parameters at higher signal-to-noise ratio, but it is strongly affected by noises. The 2nd parameter, variance of ca3, is both the most stable and the least affected by noises. Based on this analysis results, the specified discrimination parameters can be gained and are used for the discrimination of the silent/speech segments.

Two group of training samples, which are from the silent segment and the speech segment respectively, are needed while discriminating. According to the above scheme, extract discrimination parameters from these two samples and constitute two parameter sets, the silent parameter set and the speech parameter set. Extract corresponding discrimination parameters from test speech frame in the same way and calculate the distance d_s between the discrimination parameters and the silent segment set and the distance d_v between the discrimination parameters and the speech segment set. It belongs to the silent segment if $d_s<d_v$, otherwise it belongs to the speech segment. The following Mahalanobis distance is used so as to normalize the parameters.

$$d_i = (x-\bar{x}^{(i)})S_i^{-1}(x-\bar{x}^{(i)}) \qquad i=s,v \tag{14}$$

where S_i is the covariance matrix of the silent and speech segment.

4. Experiment results

In the experiment, the random noises, produced by the computer, are added to the speech segment and the above scheme is used to discriminate the silent segment from the speech segment. SNR is the same as above. The sample rate of the speech signals is 8 kHz, 8 bit quantization.

According to the parameter validity analysis, considering the computing quantity in the practical application, we tend to get the highest resolution at the least computing quantity. So the 2nd, 3rd, 5th, 8th, 11th parameter, which have the highest validity, are chosen to be discrimination parameter I, the 2nd, 3rd parameter, which have the most outstanding validity, are chosen to be discrimination parameter II. The contrast test results are shown in Table 1.

Table 1 Discrimination results of two group of parameters at different SNR

NR	20dB	15dB	10dB	5dB	0dB
Para. I	98.63%	98%	95.75%	8.57%	8.13%
Para. II	96.63%	95.75%	97%	9.5%	7.38%

From Table 1 it can be seen that, the accuracy of parameter I is lower than that of parameter II when SNR is bigger; there is no obvious difference between them when SNR is smaller, and even the discrimination effect of parameter II is better than that of parameter I. However, the computing quantity of parameter II is much less than that of parameter I, this is because parameters II are all the statistical parameters (variance and cross-zero rate). Suppose the length of the original signals is L, the length of ca3 is only $L/8$ after Mallat decomposition of 3rd order, so only the variance and cross-zero rate of $L/8$ data are needed to be calculated. Nevertheless, it's necessary to count parameters ca3, cd3, cd2, cd1, whose length are $L/8$, $L/8$, $L/4$, and $L/2$, respectively, and there is a violent increment in the computing quantity. Viewing from the practical application, we tend to use parameter II. Similarly, compared with the traditional method of counting time-domain parameters, using parameter II in wavelet domain for discrimination decreases the computing quantity, although it needs a process of wavelet transform, it can combine the speech denoising and coding. Even if wavelet transform is used independently, a fast Mallat algorithm will not increase the computing quantity evidently.

Next is an experiment separating the silent segment from a piece of practical speech signals by means of the above algorithm. The speech signals last 5 s, the sampling frequency is 8 kHz, 8 bit quantization, the practical SNR is estimated to be 17 dB. The results are shown in Figure 3.

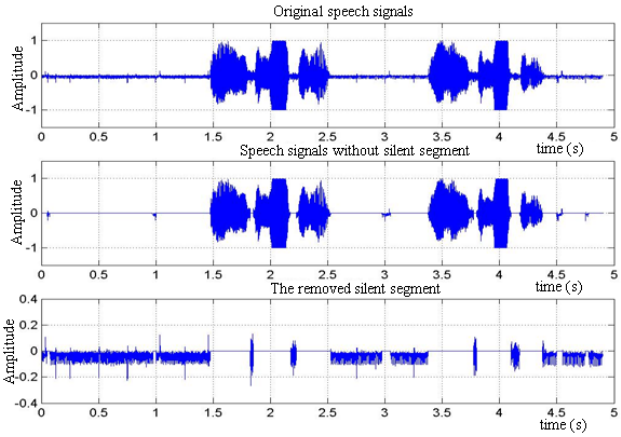

Figure 3. Experiment results of practical speech signals

5. Conclusion

In order to overcome the limitations of the traditional parameters in detecting silent/speech segments, a detection method based on wavelet transform parameters is proposed in this paper. The validity analysis method of fuzzy entropy parameters is introduced to discuss the validity of the transform parameters, and then the highest validity transform parameters are extracted to be the discrimination parameters. Taking this parameter as the discrimination, lots of contrast experiments are carried out in different noise sources and SNRs, and the discrimination accuracy has been increased evidently compared with the traditional parameters. In the end, the analysis example of eliminating the silent segment from the practical speech signals, and the results show that the parameters are effective and reliable.

Reference

[1] Zhao Li. Speech signals processing. Beijing: Machinery Industry Press, 2003.

[2] Zheng Wenming, Zhao Li, Zhou Cairong. Pattern classification based on fast KNFL method and application in quiet speech identification. Journal of Circuits and Systems, 2003, 8(2): 71-73.

[3] Baraniuk R G. Compressive sensing[J] . IEEE Signal Processing Magazine, 2007, 24(4) : 118- 121.

[4] Giacobello D, Christensen M G, et al. Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction[J] . IEEE Signal Pro-cessing Letters, 2010, 17(1) : 103- 106.

[5] Figueiredo M A T，Nowak R D，Wright S J，et al． Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems ［J］． Selected Topics in Signal Processing，2007，1 (4) : 586 597．

[6] Gemmeke J F, hamme H V, et al. Compressive sensing for missing data imputation in noise robust speech recognition[J] . IEEE-Journal of Selected Topics in Signal Processing , Special Session on Compressive Sensing, 2010, 4(2) : 272-287.

[7] Chen Sigen. Voice endpoint detection methods based on entropy function. Acoustics and Electronics Engineering, 2001, 1: 28-30.

[8] Wang Qinyun，Zhao Li，Zou Cairong． Acoustic source localization based on adaptive subgradient projection in digital hearing aids［J］． Journal of Southeast University: Natural Science Edition，2009，39 (4) : 667 672． (in Chinese)

[9] Davies M，Daudet L． Sparse audio representations using the MCLT［J］． Signal Processing，2006，86 (3) : 457-470．

[10] Xu Wang; Ding Qi; Qang Bingxi. A speech endpoint detector based on eigenspace-energy- entropy. Journal of China Institute of Communications, 2003, 24(11): 125-132.

Shubha Kadambe&G..Faye Boudreaux-Bartels. Application of the Wavelet Transform for Pitch Detection of Speech Signals. IEEE Trans. Inform. Theory, 1992, 38(2): 917-924.