Effective Speech Endpoint Detection Algorithm For Voiceprint Recognition

Yan Wang[1, a], Longfei Zhang [2, b]

[1]First Research Institute Of The Ministry Of Public Security OF PRC, Beijing, 100081, China

[2] School of Software, Beijing Institute of Technology, Beijing, 100081, China

[a]email: 18911610692@163.com, [b]email: longfeizhang@bit.edu.cn

Keywords: Voiceprint processing, Speech endpoint detection, Voiceprint recognition

Abstract. Speech voiceprint recognition with noise in complex real phone channel environment is still a critical challenge even the recognition method works well enough in non-noise situation. Background noise, especially dial tone of voice, which is the voice from surrounding disturbs the accuracy of recognition. One key problem of voiceprint processing is how to locate when the voice start and stop, and another one is how to remove all kinds of noise effectively. In this paper, we tackle these two problems and propose an endpoint detection algorithm which based on a double threshold method by processing short-time energy and linear prediction cepstrum distance. By compensating the high frequency part of the speech signal and the frequency spectrum of the signal become flat, we avoid the energy losing of small voice signal and improve the accuracy of detection. Our algorithm remains the principle of speech signal with little cost. Experiment shows the effectiveness of our algorithm both in public voiceprint dataset and real public security case dataset.

Introduction

With the rapid development of network communication and information processing technology, the amount of telephone communication home and abroad grow rapidly, which have accumulated a lot of voice data. Facing such large amount of the growing voice data, it is a huge workload to use manual work to collect the data. Moreover, it could not to find, alarm, report, and then to provide efficiently clues to the public security business application in a relatively short time.

Voiceprint recognition realized the intelligent collection of voice signal, which extracts the characteristic parameters of the speaker, and trains models to recognize the speakers. Those noises lies in voiceprint can effect the performance of recognition, such as background noise, and ringtones, etc. Voiceprint recognition is firstly to test the endpoint, segment phonetic and non-phonetic of speech signal, and then to extract and recognize the phonetic characteristics. How to carry out in the preprocessing stage of endpoint detection and removal of these noises is a difficult problem in voiceprint recognition technology.

This paper analyzes the noise characteristics under phone channel environment, studies on speech’s pretreatment technology, and optimizes model of voiceprint recognition. This paper introduces an algorithm employing short-time energy and linear prediction cepstrum distance double threshold method to reduce noise and improve the accuracy rate of recognition.

Table 1. Noise types and characteristics

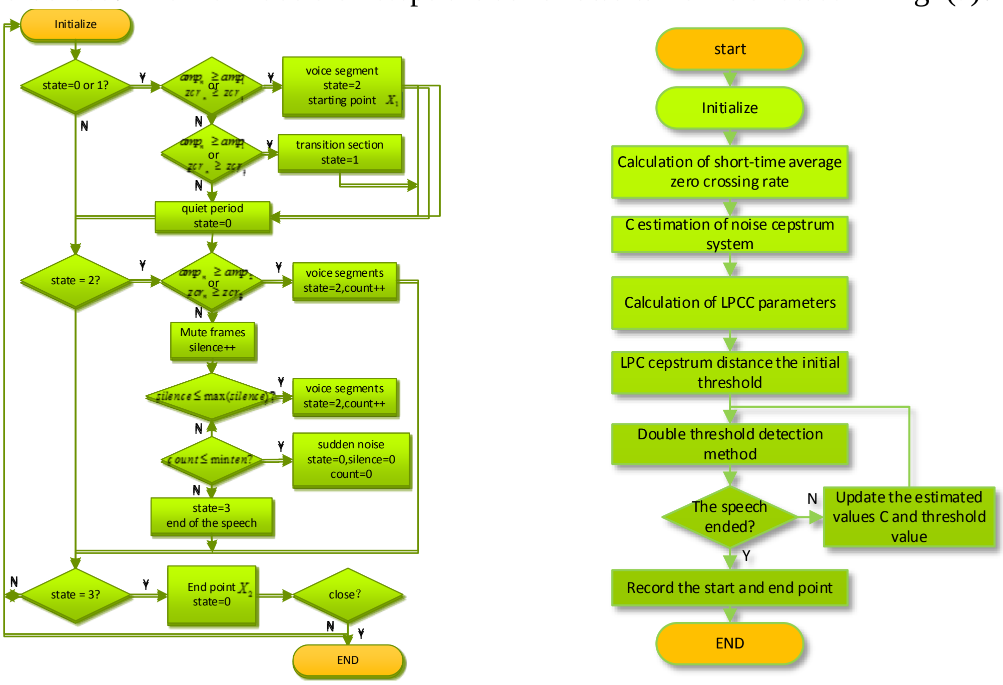
Number	Noise types	Noise Characteristics
1	The noise from people	Fill in the sound, Cough, Laughter, Breathing, Make clicks, etc.
2	Garbage voice	Color ring tone, A dial tone, Hang up the sound, Operator notes, etc.
3	Channel noise	Telephone adjustment, Telephone phonic, Square wave noise, etc.
4	Environmental noise	Hit the microphone sound,Background noise and Background voice etc.
5	Special pronunciation	Overlapping pronunciation, Short-term meaningless sounds, etc.

When application of voiceprint recognition phone channel, there has problems such as the speaker speech signal is not stable. Natural voice telephone conversations have unique noise and pronunciation characteristics. Endpoint detection is an important procedure to find the voice signal through the starting point and end point, remove noise and mute, integrate the useful speech segment to be a new voice signal.

Different sources of noise（As shown in Table 1.）

Short-term energy threshold Algorithm

Endpoint detection based on short-time energy main idea is more compared with the corresponding threshold comparison of feature parameters are extracted from the signal. When the SNR is high with short time energy, detection sound section and the silent period. Voiceless zero crossing rate high, voiced the zero crossing rate is low, with zero crossing rate to detect the unvoiced and voiced. The main decision steps are as follows which is showed in Fig1(a)：


Initialize
state=0 or 1?
Y
N
$amp_n \geq amp_1$ or $zcr_n \geq zcr_1$
voice segment state=2 starting point X_1
transition section state=1
quiet period state=0
state = 2?
$amp_n \geq amp_2$ or $zcr_n \geq zcr_2$
voice segments state=2,count++
Mute frames silence++
$silence \leq max(silence)$?
voice segments state=2,count++
$count \leq minten$?
sudden noise state=0,silence=0 count=0
state=3 end of the speech
state = 3?
End point X_2 state=0
close?
END
start
Initialize
Calculation of short-time average zero crossing rate
C estimation of noise cepstrum system
Calculation of LPCC parameters
LPC cepstrum distance the initial threshold
Double threshold detection method
The speech ended?
Update the estimated values C and threshold value
Record the start and end point
END

(a) (b)

Figure 1. (a)is the flow chart of our Short-term energy threshold algorithm and (b) is the flow chart of the linear prediction cepstrum distance threshold algorithm

The background noise, speech signal energy and the average zero crossing rate are exploited to calculate two border thresholds, respectively as amp1,amp2,zcr1, and zcr2.

Assumed that the speech signal is $\{x(m)\}$,then the short-time average zero crossing rate is：

$$Z_n = \frac{1}{2}\sum_{-\infty}^{\infty}\{sgn[x(m)] - sgn[x(m-1)]\}\cdot w(n-m) \tag{1}$$

Voice over $\{x(m)\}$ according to the sign function, determine whether the occurrence of sign changes.If there have been zero frequency, then the first-order differential operation take the absolute value, and finally through a low-pass filter to get the short-time zero crossing rate. The zero crossing rate of devoiced segments is high. The zero crossing rate of voiced segments is low. So short time average zero crossing rate can diagnose the devoiced and voiced segments. Window function w(n-m), the rectangular window, can be represented as:

$$Z_n = \frac{1}{2}\sum_{-\infty}^{\infty} sgn[x(m)] - sgn[x(m-1)] \tag{2}$$

Short-time energy can detect voiced, short-time zero crossing rate can detect devoiced. The short-time average energy and short-time zero crossing rate as the speech signal endpoint detect threshold. The speech signal can be divided into four kinds of state, such as the mute, mute to speech, speech, end of the speech. There are four main states : state∈[0,1,2,3]. The starting point and ending point of speech are stored in avector X_1 and X_2. amp_n is the short-time energy of speech signal. zcr_n is the short time average zero crossing rate of speech signal. amp_1 andzcr_1 is high threshold value .amp_2 and zcr_2 is corresponding to the low threshold.

The characteristic parameters of this method can classify between devoiced and voiced segment. The experimental results show that the voiced segment energy is significantly higher than devoiced speech segment. In this paper, by setting the energy threshold method to distinguish the voiced and unvoiced speech interval. In the case of high signal noise ratio (SNR), this method can distinguish between sound and silent as the auxiliary characteristic parameters used in voiceprint recognition.

Linear prediction cepstrum distance threshold Algorithm

Cepstrum distance feature has good stability in the complex environment. Based on the main idea of LPC cepstrum distance threshold method: voice signal sampling points are related. The speech signal can approximate the speech signal sampling value through several speech signal sampling linear combination.

Sampling point is relevant. It can approximates the speech signal sampling value through several speech signal sampling linear combination. When the error of linear prediction sampling and actual speech sampling value close to the most hours, only LPC coefficients can be determined. The transfer function H(z) of system are connected in series by radiation model and channel model. The actuating signal U(z) through the system H(z) get speech signal S(z). The convolution of system function h(n) and incentive u(n) is a time-domain speech signal. H(z) and U(z) is the product of the frequency domain of speech signal. LPC analysis of speech signal is determined by S(n) to obtain S(w) parameters.

As we all known before a frame signal cepstral vector is generated, according to this calculation method to update the environment noise cepstrum estimation C. Where C_t represents the previous frame signal cepstrum vectors, t is the previous frame signal frame number, α represents the time adjustment parameter. Based on the LPCC_D-ZCR endpoint detection using LPC cepstrum distance as the first threshold level, Set high and low threshold for $T_1$and T_2. In the short time average zero crossing rate as the level second threshold, set high and low threshold for zcr_1 and zcr_2. The update threshold and its specific end point detection steps to continuously according to environment noise, instead of using fixed threshold to determine the endpoint. Flow chart design of the algorithm is as Fig.1(b) .

Linear prediction analysis (LPC) is known as a speech signal, obtaining a set of prediction coefficients of $a_i$in a certain criterion, so that in a short speech waveform mean square prediction error value of the minimum. The short-time stationary of speech signals, first define the short-time signal $s_n(m)$ points starting point of $\varepsilon_n(m)$ and the error signal n to calculate the LPC parameters:

$$S_n(m) = s(n+m),\quad \varepsilon_n(m) = \varepsilon_n(n+m) \tag{3}$$

Let $S(w)$spectral density function for the signal energy for the logarithmic transformation, and then the inverse Fourier transform of signals from the cepstrum:

$$\log S(w) = \sum_{n=-\infty}^{\infty} C_n e^{-jnw},\quad c_0 = \int_{-x}^{x} \log S(w)\frac{dw}{2\pi} \tag{4}$$

Where C_n cepstral coefficients, is a real number, and a $c_n = c_{-n}$, and

Let $S(w)$ and $S^{'}(w)$for the spectral density function of a pair of Parsavel theorem, by knowledge, mean square cepstrum distance by the log spectral distance said.

$$d_{cep}^2 = \frac{1}{2\pi}\int_{-x}^{x} |\log S(w) - \log S^{'}(w)|^2 dw = \sum_{n=-\infty}^{\infty} (c_n - c_n^{'})^2 \tag{5}$$

Where c_n and $c_n^{'}$ 'are $S(w)$ and $S^{'}(w)$ in the cepstral coefficients. The logarithmic spectrum of mean square distance reflects the difference between the two signal spectrum to set the signal discrimination parameter. Cepstrum distance method is to set the distance threshold and signal cepstrum distance to compare, and judge each frame signal is speech or noise. The voiced speech cepstrum distance is small and short time energy, cepstral distance unvoiced large short-term energycenter, short-time energy is small and cepstrum distance of silence segment center, according to these characteristics by using the method of short-time energy and cepstrum distancein combination with the endpoint detection.

Experiments

The speech dataset in our experiment is TJS telephone dataset which is an unpublicized real world dataset and recording form Tianjin public security cases. Telephone recording data code conversion into a sampling rate of 8KHz, PCM voice 8Bit accuracy, obtained a about 100 a speech samples, each speech sample length of 60 seconds. The following processing of a single voice data:

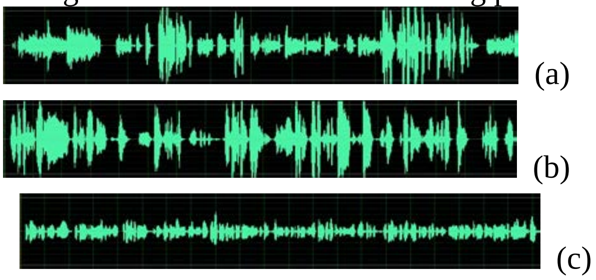


Figure 2. (a) is the original speech waveform, (b) is Short-time energy threshold method proceed speech waveform and (c) is Linear prediction cepstrum distance threshold method proceed speech waveform

One situation can be seen from the figure, the algorithm can detect and remove lower energy section of the speech signal effectively, and voice data obtained from the speech waveform after endpoint detection is complete.

The double threshold endpoint detection algorithm if can replace the artificial way to deal with the speech samples, artificial cost is greatly reduced. The experimental test, we used the artificial processing method with double threshold method to dispose sampling voice data, and then call the voiceprint recognition engine to identify the speaker, statistical results are shown in Table 2.

Table 2. Voiceprint recognition results by using different algorithms

The endpoint detection method	Artificial processing sample method	Double threshold method
System identification	92.34%	87.96%

From experimental results, we can find that the first double threshold method is used to sample pretreatment all voice files, and key person's speech samples for artificial processing, can greatly save police sample processing time. Tianjin development and deployment of the voice print library management information system, utilizing the double threshold endpoint detection algorithm in the voice print library system, good noise reduction effect were obtained.

Conclusion

This paper designed a short-time energy and linear prediction cepstrum distance double threshold method, which can effectively reduce the end point judgment tail long occurrence. The voice activity detection algorithm of double threshold method for pretreatment of voice data in real cases, experiments show that, under the low SNR environment, the algorithm reduces the amount of signal transmission and recognition computation load, which is able to better distinguish voice and noise, has better robustness and higher detection accuracy rate.

Based on short-time energy and linear prediction cepstrum distance application endpoint detection algorithm of double threshold method voiceprint database management information system in Tianjin, it reduced the workload of manual processing voice sample, improved the quality of voiceprint database sample, effectively reduced the speech distortion, and improved the recognition of voiceprint recognition rate.

Acknowledgement

This material is based upon work supported by the Key Technologies Research and Development Program of China Foundation under Grants No. 2011BAK05B00 and 012BAH38F02-04.

Reference

[1]Tan Z H,Lindberg B.Advanced in Pattern Recognition-Automatic Speech Recognition on Mobile Devices and over Communication Networks. London:Springer-Verlog London Limited,2008.4,4-16
[2]Janet M B,Li D,James G,et al.Research Developments and Directions in Speech Recognition and Understanding,Part1.IEEE Signal Processing Magazine,2009,26(5):32-45
[3]Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models.By Douglas A.Reynolds,Member,IEEE,and Richard.Rose,Member,IEEE,2010,7-9
[4]J.M.Gorriz,J.Ramirez,E.W.Lang,C.G.Puntonet.Hard C-Means Clustering for Voice Activity Detection Speech Communication,2006,48(12):1624-1637
[5]Liang Y,Liu X,Lou Y,Shan B.An improved noise-robust voice activity detector based on hidden semi-markov models.Pattern Recognition Letters,2011,32(7):1034-1045
[6]Shin J W,Chang J H,Kim N S.Voice Activity detection based on statistical models and machine learning approaches.Computer Speech and Language,2010,24(3):525-536
[7]Huang H Y,Lin F H.A speech feature extraction method using complexity measure for voice activity detection in WGN.Speech Communication,2009,51(9):724-736