ISSN : 0974 - 7435 Volume 10 Issue 20

2014

BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 10(20), 2014 [11852-11858]

Research on speech recognition system in complex environments

Li Min[1] *, Xie Quanying [2]
[1] Electronics and Information Engineering, Chongqing Radio and TV University, Chongqing 400052, (CHINA)
[2] College of Preparatory Education, Southwest University for Nationalities, Chengdu 610041, (CHINA)
Email: lymym@126.com

ABSTRACT

As for the technology progress and the development of manufacturing technique, smart devices are applied in more and more places. The usage of smart devices reduces the workload on people greatly. Besides, it expands people's ability when work in extreme circumstances. This thesis will study the speech recognition in complex environment and acquire the effective voice message quickly in complex environment with plenty of interfering noise. Currently, most of the speech recognition products can produce voice recognition with little noise. And only in this way, they can acquire effective voice message. However, all these devices can not meet our requirements in normal environment. The speech recognition rate will drop sharp or lose efficacy in a car, which makes the current products have little adaptable range. This thesis raise a speech enhancement method who is combined with PUM model, which could eliminate the influence of noise on the speech recognition system effectively. This model improves significant recognition rate of the speech recognition system in noisy environment. It can make the speech recognition system maintain high speech recognition accuracy in vehicle's high noisy environment. By using the above technology, the speech recognition ability and the applicable scope was greatly improved. Besides, it also gives higher speech recognition accuracy than before in the same environment.

KEYWORDS

Access control mechanisms; Open system; Dynamic authorization; Cross-domain access; Centralized management.

INTRODUCTION

The speech recognition technology in our research is the latest technology with interdisciplinary feature. It is the key interface technology of current human-computer voice interaction. It is the dream of talking to computer and letting it known what you want to say. People try to find efficient methods to communicate with computer at its first appearance. Speech recognition technology appears right after that, and it gets great improvement with our persistence for decades. The speech recognition technology could be explains as follow. Smart devices recognize all input voice from which they could analyze and understand people's intention. And then they transform all the voice signals into logic information which could be recognized by computer.

The usage of computers' speech recognition and voice control provides great convenience for people's input by keyboard and mouse, especially those who have disability on hands. They can not finish input by keyboards and mouses. Because of this, It is of great necessary for us to develop a computer-based speech recognition, voice-activated system.

Currently, the accuracy of the small vocabulary with non-specific speech recognition system has been more than 90%, and it could be better among specific vocabulary. The basic speech recognition ability can meet some of the requirements for voice communication applications. For now, in some developed countries from the west of world, several speech recognition products have been put into market to do exploration on voice recognition service. However, this service is mainly used to perform application services, such as Automatically Pick-up Service, Intelligent Voice Dialing Service, Voice Smart Toys and Voice Notepad. Limited telephone network voice self-services could be performed at present. Applications which could identify and check voice messages by standard voice input have been promoted with small-scale. Statistics from these smart speech services has indicates high satisfaction among all users. The vast majority of application users obtained good voice analysis service, and they expressed high satisfaction. Therefore, we may predict audaciously that intelligent speech recognition systems would be widely used in much more fields. On this condition, research on speech recognition technology based on various environments, as a very challenging subject, will certainly benefit a lot.

BASE MODEL OF SPEECH RECOGNITION

Foundational Model of Voice Recognition

The ultimate goal we have been pursuing is to communicate with computer in natural language and the computer receives instructions meanwhile, letting the computer know what people want it to do[1]. This technology encompasses several advanced research technology includes computer bionics, computational linguistics, artificial intelligence and so on. It is high technology who translate video signal into normal binary code which could be understand by computer. Previous works in this field contribute a lot to the current base model on speech recognition. The process of speech recognition model is shown in Figure 1.

Conclusion from figure 1 indicates that the speech recognition processing model consists of four components: voice endpoint detection unit, voice feature extraction unit, speech database training and anti-noise processing unit.

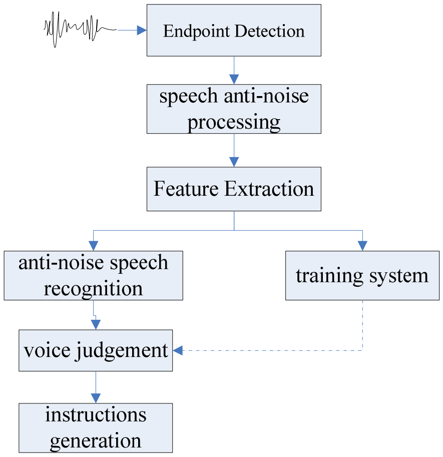

Figure 1 : Process of speech recognition processing model

Endpoint Detection

Void endpoint detection is the technology which focuses on distinguishing whether the input signal is voice. Voice breakpoint detection function is used to complete the detection of voice input. It is the foundation of the speech recognition system for voice processing. The effectiveness, as the key node in the system, is directly related to detection performance of the speech recognition system. Therefore, the breakpoint detection plays an important role in the whole processing system. There are lots of parameters during all process such as speech analysis, voice filtering and speech enhancement. All the parameters' calculation is depended on the corresponding input signal segment. In that case, only with accurate speech endpoint signal, we can perform voice processing correctly.

We make the products of FM (Frame Energy) and FZCR (Frame Zero-Crossing Rate, the number of zero-zero energy value in short-term) as the indicator of endpoint detection. The product is named ZFE (Zero Frame Energy). It represents the voice amplitude summation of this speech sample frame in a short time interval. Its computational formula lists as formula (1). We assume that there are N samples in a frame. I ($i\in[0,N]$)is one of the samples. And S[i] implies the FM of sample i. Therefore, the formula of FZCR is as shown in formula (2). In this formula, S[i-1] is the former FM of S[i]. There are also circumstances that FM is little while FZCR is much larger or that FM is large while FZCR is little on the syllable. However, the product of FZCR and FM could be maintained its stability. The value is much larger than that of sound without discourse, ensuring the reliability of speech acquisition.

$$Power = \sum_{i=1}^{N} S^{2}[i]) \tag{1}$$

$$Zero = \sum_{\substack{i=1 \\ S[i]*S[i-1]<0}}^{N} 1 \tag{2}$$

Noise statistics

We regard it as the end of effective speech when successive frames appear to be lower than the zero FM threshold after effective voice frames appear during our endpoint detection. Through this mechanism we can detect and judge the intermediate intervals of received voice, distinguishing the former and after word with better recognition accuracy. Meanwhile, the detection may weed out the superfluous or invalid voice sample values by removal of the word-word segmentation and judgments of fore-aft voice sentence. The threshold during consistence detection is concluded from noise statistics.

Noise statistics is used to determine the TSH[2,3]. And the threshold is directly related to the effectiveness of detection. On this condition, the result of noise statistics is the direct element that has influence to threshold. The threshold is the key indicator judging the beginning and ending of effective voice. The calculation formula of this threshold is as formula (3).

$$THS = \frac{1}{N}\sum_{i=1}^{N} Power\ [i]\times Zero\ [i])\times \mathrm{k} \tag{3}$$

In this formula, N means the total frames for threshold estimation. And i is the current frame who is in used. And k is a preliminary defined constant, its value is generally selected for range[2,3].

FEATURE EXTRACTION

After receiving voice input, we need to extract all the characteristic parameters. The extraction includes two important procedure: effective detection of voice signal and the compression of the recognized message. The extraction, as a processing technology, is of significant importance to speech recognition. After sampling, recognition, analysis and feature extraction, system will compare and recognize speech input that of relatively stability by comparison of Characteristics, identifying their practical significance of input voice rapidly and stably. During the extraction, we used the short-time Fourier analysis method [4]and MFCC speech feature analysis.

Short-time fourier analysis

Fourier analysis is the standard Fourier analysis on various status of random signal, such as transient, period, or balance. However, the input voice, unlike the innocent friction sound and continuous original sound input which is stable, may emerge complex speech wave with the change of time. Now, we could analysis the speech wave in short time to judge the stability. Short-time analysis is an effective way to steady the input speech wave. Short-time Fourier analysis is to distribute a long period analysis into number of fixed-length short period, reducing circumstances that could not analyze the unstable and long-period waves. In that case, we could analyze the transient variation with short-time Fourier analysis in every short period. Speech spectrum characteristic wave in a long period will take its shape after combining all of the short-period ones [5]. And its formula shows like formula (4).

$$X_{n}(e^{j\omega}) = \sum_{m=-\infty}^{+\infty} w(n-m)s(m)e^{-j\omega m} \tag{4}$$

Parameter n is the signal staging area on a specific time index. And w (n-m) is the time sequence “window” on this stage. S (m) is the original input speech sequence segment. The wider time window is, the lower time resolution is. And the frequency resolution is higher too. From this point, the longer analysis period is, the better frequency resolution is. But this will cause much longer analysis period than expected. It does not meet the requirements of our short-term analysis. Besides, we could not achieve the short-term feature extraction goal. On this condition, we need to choose appropriate time window and its coverage area. We choose the time window with moderate size. We prefer the time window with moderate size, and the window shape with high frequency resolution and smaller sidelobe**[6]**. Through the comprehensive analysis, Hamming window was chosen. Function of Hamming time window is formula (5).

$$\begin{cases} 0.54-0.46\times\cos\left(\dfrac{2\pi n}{N-1}\right), 0\le n\le N-1 \\ w(n)=0 \end{cases} \qquad \textbf{(5)}$$

N is the value of window bandwidth.

Analysis of voice MFCC

Analysis of MFCC**[7]** is the human-like analysis which accepts voice from outside. We considered the linear relationship between waves’ peaks-troughs and its frequency. By the bionic way, computer would get voice information which is most like information gotten by our ears. This receive features match the simulation value with auditory feature of Mel frequency size better. In this system, we distributed a voice frequency band into a series of triangular filter voice band sequence according to the Critical band, forming the Mel filtering group.

After receiving weighted summation of all signal amplitude from Mel filter frequency bandwidth, we make it as the bandpass filter output. And then, all filter outputs do logarithm, a further DCT will contribute MFCC. Bandpass filter output calculate from input voice amplitude spectrum $|X_n(k)|$. Formula (6) lists the details.

$$m(l)=\sum_{k=o(l)}^{h(l)} W_l(k)\,|X_n(k)| \qquad l=1,2...,L \qquad \textbf{(6)}$$

After receiving weighted summation of all signal amplitude from Mel filter frequency bandwidth, we make it as the bandpass filter output. And then, all filter outputs do logarithm, a further DCT will contribute MFCC. Formula (7) lists the process of transformation.

$$c_{mfcc}(i)=\sqrt{\frac{2}{N}}\sum_{l=1}^{L}\log m(l)\cos\{(l-1/2)\frac{i\pi}{L}\} \qquad \textbf{(7)}$$

Voice Training and Recognition

VOICE TRAINING

$Baum\text{–}Welch$**[8]**, the solution of the parameter estimation algorithm in Hidden Markov model, was used during our voice training. It chooses an observation sequence $0 = o_1, o_2, \cdots, o_T$ among the parameters. The algorithm will determine a model parameter *M={A,B,Π}* to make sure that *P(O|M)* is the largest. It is a functional extremum problem to make sure that *P(O|M)* is the largest by variation of M. Algorithm $Baum\text{–}Welch$ use recursion extremum to partly amplify $P(O/M)$ in observation sequence step by step, obtaining the optimized model parameter *M={A,B,Π}*. By training the input voice, we may get optimized training model. And it will be used in subsequent speech recognition. The detail steps are: (1) Choose a section of voice signal $0 = o_1, o_2, \cdots, o_T$ as the observation sequence and model parameter *M = {A, B, Π}*. (2) Take sequence $0 = o_1, o_2, \cdots, o_T$ from head to tail. (3) Probability that transform *Si* to *Sj* at time *t* is γ*t(i,j)*,transform probability is calculated by formula (8).

$$\gamma_t(i,j)=\frac{\alpha_{t-1}(i)a_{ij}b_{ij}(o_t)\beta_t(j)}{a_T(N)}=\frac{a_{t-1(i)}a_{ij}b_{ij}(o_t)\beta_t(j)}{\sum_i a_i(i)\beta_i(i)} \qquad \textbf{(8)}$$

Meanwhile, we could acquire the probability that sequence has the status of Si at time t, just as formula (9) shows.

$$\sum_{j=1}^{N}\gamma_t(i,\ j)=\frac{a_t(i)\beta_t(i)}{\sum_i a_i(i)\beta_t(i)} \qquad \textbf{(9)}$$

Therefore, the expected times of sequence s_i transforming into S_j is $\sum_t \gamma_t(i, j)$; and the expected times of sequence S_i transforming out is $\sum_j \sum_t \gamma_t(i, j)$. Revaluation formulas with export model are shows in formula (10) and formula (11). Then we could get a group of new parameters to generate new models.

$$\hat{a}_{ij} = \frac{\sum_t \gamma_t(i, j)}{\sum_j \sum_t \gamma_t(i, j)} = \frac{\sum_t a_{t-1(i)} a_{ij} b_{ij}(o_t) \beta_t(j)}{\sum_t a_t(i) \beta_t(i)} \tag{10}$$

$$\hat{b}_{ij}(k) = \frac{\sum_{t:0_t=k} \gamma_t(i, j)}{\sum_t \gamma_t(i, j)} = \frac{\sum_{t:o_t=k} a_{t-1(i)} a_{ij} b_{ij}(o_t) \beta_t(j)}{\sum_t a_{t-1(i)} a_{ij} b_{ij}(o_t) \beta_t(j)} \tag{11}$$

Speech Recognition

Algorithm Viterbi, as the voice input recognition algorithm in this system, is described as follows: (1) Create an array $a'_t(j)$ for every status, the original status S1 corresponds to $a'_0(1)$. Then we initialize it as 1. Others are initialized to be 0. (2) Calculate the state value and write it into $a'_t(j)$ by the symbol sequence at time t. $a_{ij} = 0$ when the status value do not change. (3) Create a new array of State record to save i which facilitate the maximum $a'_t(j)$. (4) Output the array when there is status change. The status array we get is the best status sequence, and it is also the best recognition of input voice.

ANTI-NOISE IMPROVEMENT AND BUILDING NEW MODEL IN COMPLEX ENVIRONMENT

Anti-noise improvement

We bring in several anti-noise recognition technologies in our system for better speech recognition accuracy in complex environment. All these technologies enhanced the application possibility of the speech recognition system in practical environment. The existing popular speech anti-noise technologies includes: noise compensation method[9], speech extension method[10], feature removal anti-noise method, noise extraction and speech-insensitive method. Every method has its certain circumstance, but they are not the ideal way to handle complex environment[11]. Therefore, we optimized our anti-noise algorithms in this system.

Our algorithm used in this system is the optimization based on characteristic-abandon algorithm and speech enhancement algorithm. The main point is to filter the speech wave which contains noise with the help of speech enhancement algorithm, eliminating the broadband noise effectively. And then, we abandon the son-band speech characteristics which have been polluted by noise with the help of characteristic-abandon algorithm, leaving pure son-ban speech characteristics[12,13]. The algorithm provides characteristic-abandon model speech fragments which is polluted by noise with partly voice message. In this way, characteristic-abandon method could be applied to speech signals on the full spectrum band which were polluted by noise.

However, this optimization has also defects: (1) Speech enhancement algorithm could clear noise which could be estimated well, but it may do less when is confronted with time-changing noise and noise which could not be estimated. (2) Speech enhancement may cause the distortion of input speech signals, introducing new noise signals when de-noising. In that case, the noise information would be more complex and more difficult to eliminate effectively. Since their different advantages, scenes and complementary anti-noise process between the two algorithms, their combination may be better way. The combination will inherit advantages and abandon the defects of the two algorithms.

We improved the repeat speech enhancement method of Wiener filtering[14] during speech enhancement application. Traditional repeat speech enhancement method of Wiener filtering was used to speech enhancement on plus-noise condition. The noise model could be illustrated with formula (12）.

$$x(t) = s(t) + n(t) \tag{12}$$

In this formula, x(t) is the speech signal of noise. S(t) is the clean signal without noise. And n(t) is the full noise signal in speech[15]. The result of filtering noise when processed with traditional repeat speech enhancement method of Wiener filtering is not quite clear. So we improved the traditional repeating Wiener filter. The main optimization is to use the linear prediction analysis circularly to enhance the speech. To reducing workload when processing cycle forecast analysis, we proposed a new promotion. In this promotion, the amount of calculation is much smaller and the enhancement works well. The points which were taking into consideration list as follows. (1) Subtract $\alpha \hat{P}_n(\omega)(a > 1)$ when the frame has high MMSE-STSA. It will provide protruding speech spectrum, less pure tone noise and better performance on noise reduction. (2) The main reason which cause spectrum' distortion after filtering is the difference between noise power spectrum $P_n(\omega)$ and its estimated value $\hat{P}_n(\omega)$. Finding the average among several frames on voice signals spectrum with noise will reduce distortion

after filtering. The initialization improvement formula and repeating filter improvement formula is formula (13) and formula (14).
Initialization improvement formula:

$$\begin{cases}\hat{P}_s(\omega)_0=P_x(\omega)-\eta\hat{P}_n(\omega),\ P_x(\omega)-\eta\hat{P}_n(\omega)\geq\mu P_x(\omega)\\ \hat{P}_s(\omega)_0=\mu P_x(\omega),\ P_x(\omega)-\eta\hat{P}_n(\omega)<\mu P_x(\omega)\end{cases}\qquad\textbf{(13)}$$

Repeating filter improvement formula:

$$\begin{cases}H(\omega)_i=\dfrac{\hat{P}_s(\omega)_i}{\hat{p}_s(\omega)_i+\hat{P}_n(\omega)},\ \hat{P}_s(\omega)_{\mathrm{i}}-\lambda\hat{P}_n(\omega)\geq\psi P_x(\omega)\\ H(\omega)_i=\dfrac{\psi P_x(\omega)}{\psi P_x(\omega)+\hat{P}_n(\omega)},\ \hat{P}_s(\omega)_{\mathrm{i}}-\lambda\hat{P}_n(\omega)<\psi P_x(\omega)\end{cases}\qquad\textbf{(14)}$$

The effect is significant through above optimization. The speech signals we acquire finally only contain some local-band noise signal interference.

New Speech Model

In our new model, we combine two complementary
models: the optimized repeating Wiener filter model and Characteristics abandon model. During the speech recognition, the new model will firstly use the repeating Wiener filter to filter the speech with noise. After receiving son-ban MFCC speech characteristics, we take it as input of characteristic-abandon algorithm model. Finally, we recognize speech with the characteristic-abandon algorithm model. New model construction is illustrated in figure 2.

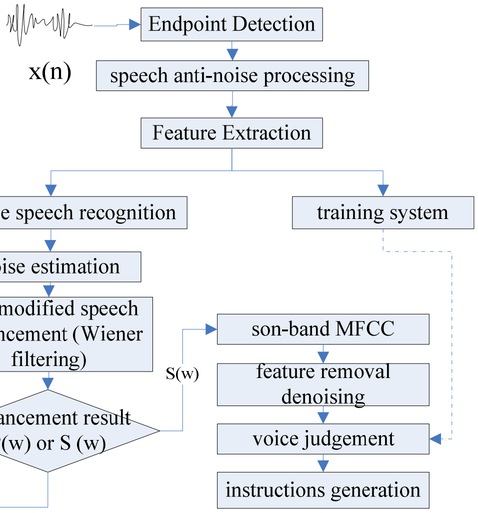
x(n)
Endpoint Detection
speech anti-noise processing
Feature Extraction
anti-noise speech recognition
training system
noise estimation
The modified speech enhancement (Wiener filtering)
enhancement result P(w) or S (w)
S(w)
P(w)
son-band MFCC
feature removal denoising
voice judgement
instructions generation

Figure 2 : New model of speech recognition and processing

The usage of the combination of two complementary models: the optimized repeating Wiener filter model and Characteristics abandon model, contribute us to acquire useful information in speech wave, achieving speech recognition. Figure 3 shows the original speech spectrum. Figure 4 is the speech spectrum after reducing noise.

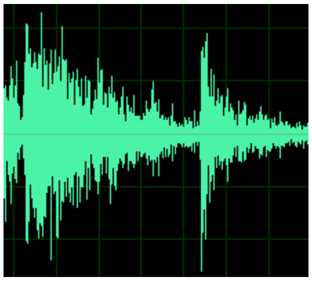

Figure 3 : Original spectrum

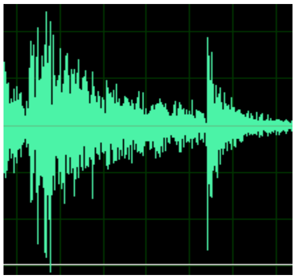

Figure 4 : Spectrum after de-noising

CONCLUSIONS

In this subject, we did in-depth study focuses on speech recognition technology. We proceeded our research mainly focus on speech recognition technology which could work well in complex environment on the condition of existing speech recognition systems. We were working on strengthen the man-machine interaction ability in a complex environment, expecting broader areas that speech recognition technology could be applied to. The existing speech recognition model consists of four components: voice endpoint detection unit, voice feature extraction unit, speech database training and anti-noise processing unit.

This paper presents better capability on anti-noise and noise reduction of speech recognition system based on former works. The improvements make it possible to recognize voice in noisy outdoor environment. The discarding characteristics method and the speech enhancement method are combined in our anti-noise optimization. The model of speech enhancement and discarding characteristics method, as a new solution, is really effective to eliminate the influence of noise on the speech recognition system, improving the recognition accuracy in a noisy environment. It also provides higher speech recognition accuracy in noisy car environments. The above techniques, what we do in this thesis, improved the voice recognition capabilities and scope greatly, as well as the accuracy of speech recognition.

REFERENCES

[1] Han Wen-Jing, Li Hai-Feng, Ruan Hua-Bin, Ma Lin; Review on Speech Emotion Recognition [J], Journal of Software, **25(1)**, 37-50 **(2014).**

[2] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, Shrikanth Narayanan; Emotion recognition using a hierarchical binary decision tree approach[J]. Speech Communication, **9, (2011).**

[3] Jiang Jianzhong, Zhang Dongfang, Zhang Lianhai; Speech enhancement algorithm for high noise environment[J], Computer Engineering and Applications., **6** , **(2012).**

[4] Moataz El Ayadi, Mohamed S.Kamel,Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases[J], Pattern Recognition. **3**, **(2010)**.

[5] LI Yin-Guo, OU YANG Xi-Zi, ZHENG Fang. Noise Robustness Analysis With Auditory Feature During Speech Recognition[J], Journal of Tsinghua University (Science and Technology)., **53(8)**, 1082-1086, **(2013).**

[6] Hu Zhengquan, Zeng Yuming, Zong Yuan; E.G. Improvement of Mfcc Parameters Extraction In Speaker Recognition[J], Computer Engineering and Application., **50(7)**, 217-220 **(2014).**

[7] Pang Cheng, Li Xiaofei, Liu Hong. Speaker Gender Recognition Based On Combining The Contribution Of Mfcc And pitch features[J], Journal of Huazhong University of Science and Technology (Nature Science Edition)., **41, (2013).**

[8] Siqing Wu,Tiago H.Falk,Wai-Yip Chan; Automatic speech emotion recognition using modulation spectral features[J], Speech Communication, **5, (2010).**

[9] Zhou Yuehai, Li Fanglan, Tong Feng; The microphone array speech enhancement and HMM recognition joint processing in noise environment [J], NCMMSC'2013. **(2013).**

[10] Ning Xiangyan, Jing Hao; Application of twice decision fusion system to speaker recognition [J], Engineering Journal of Wuhan University. **4, (2011).**

[11] He Ling, Yuan Ya-Nan, Yin Heng; e.g. Automatic Hypernasal Detection Based on Acoustic Analysis in Cleft Palate Speech [J], Journal of Sichuan University (Engineering Science Edition), **2, (2014).**

[12] Max A.Little, A.E.Declan Costello, Meredydd L.Harries; Objective Dysphonia Quantification in Vocal Fold Paralysis: Comparing Nonlinear With Classical Measures[J], Journal of Voice, **1, (2011).**

[13] Liu Hai-Bo, Li Hui, Ling Zhen-Hua; The research on pitch extraction method for voice activity detectionbased on periodic decomposition[J], Journal of University of Science & Technology China, **2, (2012).**

[14] Lee Yun-Kyung,Kwon Oh-Wook; Application of shape anal-ysis techniques for improved CASA-based speech separation, IEEE Transactions on Consumer Electronics. **(2009).**

[15] Gao Liu-yang, Zhu Wen, Sang Zhen-xia, Mi Lan; Speech enhancement algorithm based on improved spectral subtraction[J]. modern electronics technique. **17, (2012).**