Ekoloji 28(107): 1561-1565 (2019)

Speech Recognition in High Noise Environment

Chunling Tang [1*], Min Li [1]

[1] College of Electronic and Information Engineering, Chongqing Technology and Business Institute, Chongqing 401520, CHINA

[*] Corresponding author: 18393135@qq.com

Abstract

Most of the existing speech recognition products require the speech to be extracted in a small noise or no noise environment. It aims to obtain the correct speech information. However, such speech recognition cannot be used in our normal environment. For example, the recognition rate of the voice is drastically reduced in running cars and trains. Sometimes the voice cannot be recognized at all. Therefore, the adaptability range of the existing voice product becomes small or not applicable. This study aims to investigate the speech recognition in high noise environment. We proposed a new method, which used speech enhancement combined with discard feature model. The new method can effectively eliminate the influence of noise on the speech recognition system and obtain a complex environment with a large amount of interference noise. In this case, the correct voice information in the voice information is quickly identified. The recognition rate of the speech recognition system is improved in the high noise environment of automobile.

Keywords: speech recognition, endpoint detection, feature extraction, voice training, anti-noise improvement

Tang C, Li M (2019) Speech Recognition in High Noise Environment. Ekoloji 28(107): 1561-1565.

INTRODUCTION

Speech recognition makes car driving more fun such as using voice navigation or finding a song in MP3. Speech recognition has been widely used in the automotive field. However, the high-speed driving of cars and the noise generated in complex environments have severely affected the recognition of speech.

Speech recognition technology is a widely used computer application technology. The applications include voice dialing, voice navigation, indoor device control, voice document retrieval, and simple dictation data entry. It recognizes the human language input through the smart device, analyzes and understands people's intentions, and then transforms the voice signal in the process into the corresponding logical information that can be recognized by the computer. Speech recognition technology combined with other natural language processing techniques such as machine translation and speech synthesis technology can be used to build more complex applications, such as speech-to-speech translation. Speech recognition technology is now widely used in car navigation. The topic of this study is speech recognition in high-noise environments in automobiles.

BASE MODEL OF SPEECH RECOGNITION

Speech recognition, also known as Automatic Speech Recognition (ASR), aims to convert vocabulary content in human speech into computer readable input such as buttons, binary codes or sequences of characters. 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.

A commonly used speech recognition method is to pre-process and extract the speech signal to establish a template library, and to match the to-be-recognized speech signal with the template library according to the speech feature and perform speech recognition according to the matching distance. A typical speech recognition process (Lee et al. 2009) includes basic units such as voice preprocessing, feature extraction, training, and recognition. The process of speech recognition model is shown in **Fig. 1**.

SIGNAL PREPROCESSING

The voiced preprocessing is an important basis for speech recognition. The preprocessing of the speech signal includes: pre-filtering, digitization of the speech signal, pre-emphasis, windowing and framing, noise suppression and endpoint detection. Endpoint detection is a detection technique that detects an input

 Received: 6 Dec 2017 / Accepted: 17 Sep 2018

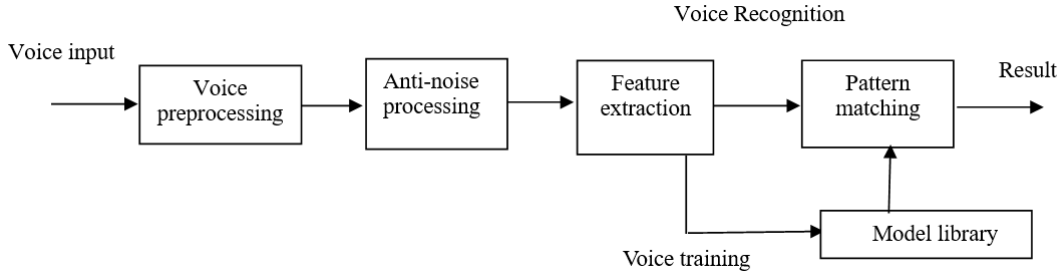

Fig. 1. Elements diagram of speech recognition

signal and determines whether the input signal is speech. It is the basis for the speech processing of the entire speech recognition system. It is a key node in the whole speech processing system, which is used to complete the function of voice input detection. Its effectiveness on voice detection is directly related to the detection performance of the speech recognition 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.

FM (Frame Energy) and FZCR (Frame Zero-Crossing Rate, the number of zero-zero energy value in short-term) as made the indicator of endpoint detection. The all is named ZFE (Zero Frame Energy). It represents the short-term speech amplitude and the time interval of the frame of the speech sample.

The formula of FZCR is as formula (1). S [i] is the latter value of S [i-1].The formula of FM is as the formula (2). We set the total number of sample values sampled in one frame to be N. i(i ∈ [0,N])is a sample belonging to it. And S[i] is the FM value of the detected i-th sample. The product of FZCR and FM could be maintained its stability.

Noise statistics are used to determine the threshold TSH (Jiang et al. 2012). The calculation formula as formula (3). This threshold is related to the validity of the endpoint detection. In formula (3), i is the ith frame used in the current calculation, k is a predefined constant parameter value, N is the total number of frames used for threshold estimation.

$$Zero = \sum_{\substack{i=1 \\ S[i]*S[i-1]<0}}^{N} 1 \qquad (1)$$

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

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

FEATURE EXTRACTION

Feature extraction is a very important processing technique for speech recognition. During the extraction, we used the short-time Fourier analysis method (Moataz et al. 2011) and MFCC (Hu et al. 2014) speech feature analysis usually.

Fourier analysis refers to the standard Fourier analysis of information numbers for random signals in the case of transients, cycles or balances. However, the input speech will appear complex speech waves over time, rather than the simple fricative or continuous input of the original sound. At present, it can be judged by analyzing whether the speech wave of the speech remains stable in a short time, and such short-term analysis is an effective processing method for solving the instability of the input speech. Short-time Fourier analysis is the analysis of a long-period analysis into a short period of fixed length. In this way, it is possible to avoid the inability to analyze unstable long-period problems, so that each transient change of the signal in each short cycle is analyzed according to a standard Fourier transform. All information transient short-period consolidated into long-period spectral characteristics of speech.

For the time window, the wider the time window, the lower the time resolution and the higher the frequency resolution. From this point, it is seen that there is better frequency resolution in the longer analysis period, but this will cause the analysis period. Too long, does not meet our short-term analysis requirements, cannot achieve the purpose of short-term feature extraction. We generally choose a time window that is moderate, and the window shape requires a frequency resolution that is as high as possible, but with a smaller side window shape. A comprehensive analysis of the Hamming window is more suitable for the analysis of input speech signals.

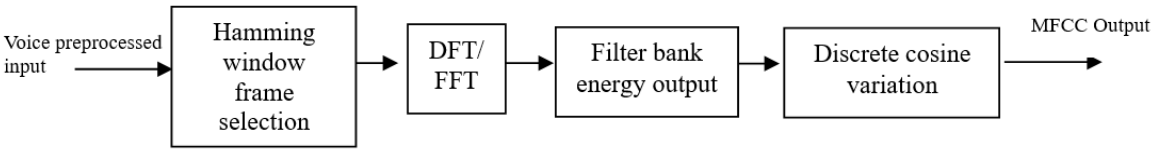

Fig. 2. Computation flow of MFCC coefficients

The MFCC speech feature analysis is based on the way the human ear accepts the external voice to design the sound collection method of the computer system through the linear relationship between the height of the sound and the frequency. Through such a bionic mode simulation system, the voice information closest to the human ear is received for analysis. This reception characteristic is more in line with the analog value of the auditory characteristic of the Mel frequency size. The calculation process of MFCC coefficient is shown in the **Fig. 2**.

In the system, we divide a set of voice bands into a series of triangular filter voice band sequences according to the critical band to form a mel filter group.

VOICE TRAINING AND RECOGNITION

The pre-processed speech signal is extracted by frame to form a series of feature vector sequences, and these feature vectors are formed. The sequence can be used for speech training and recognition.

In recent years, the hidden Markov model (HMM) (Wu et al. 2009) based on the Baum-welch algorithm has become one of the main models for the determination of the driving state of traffic vehicles. Hidden Markov process is a double stochastic process: a system used to describe short-term stationary segments of non-stationary signals. Another heavy stochastic process describes how each short-term stationary segment transitions to the next short-term stationary segment, i.e. the dynamic characteristics of short-term statistical features. Based on these two stochastic processes, the HMM can effectively solve the problem of how to identify short-term stationary signal segments with different parameters and how to track the conversion between them.

Algorithm Baum-Welch for solving parameter estimates of hidden Markov models in speech training. It maximizes P by determining an M model parameter by the algorithm by selecting an observation sequence A in the parameter values. It is determined by M that the maximum value of P is a functional extreme value problem. Algorithm Baum-Welch uses a recursive stepwise extremum to locally amplify P in the observation sequence, and finally obtains an optimized model parameter M. Thereby, an optimized training model of the speech is obtained by training the input speech for the recognition of the subsequent input speech.

The algorithm Viterbi is used as the speech input recognition algorithm in the system and is described as follows: First create an array $a'_t(j)$ for each state of the training. The initial state S1 corresponds to an array variable of $a'_0(1)$, initialized to 1, and other state values initialized to 0; The state value at this moment is calculated by the symbol sequence o_t at time t and written into the array $a'_t(j)$. $a_{ij} = 0$ if there is no transition in the state; Create an array of state records, use this array $a'_t(j)$. to save each time making the largest state I; Finally, the state array is output when the state transitions, which is the best state sequence to be obtained. This is the best recognition of the input voice of the environment.

ANTI-NOISE IMPROVEMENT

In order to improve the recognition rate of speech information in complex environment such as in automobiles, many anti-noise recognition techniques have been introduced into speech recognition system. The existing popular speech anti-noise technologies includes: noise compensation method (Zhou et al. 2013), speech extension method, 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. Therefore, we optimized our anti-noise algorithms in this system.

Algorithm that we use in this system is based on characteristic-abandon algorithm and speech enhancement algorithm. With a focus on speech enhancement algorithms for voice wave includes noise filtering, the effective elimination of broadband noise. Then, we gave up with voice features to give up the child with the help of noise pollution algorithm in the feature, leaving pure son-ban speech characteristics. The algorithm provides a speech piece model feature abandoned, which is part of the voice message of noise pollution. Thus, the method features may be applied to

abandon the whole band speech signal contaminated by noise. 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 have mainly improved the repeated Wiener filtering speech enhancement method in the application of speech enhancement method. We use a Wiener filtering algorithm based on a priori SNR estimation proposed by Scalart et al. For traditional repeater Wiener filters are often used for speech enhancement under additive noise conditions. The noise model is shown in formula (4).

$$x(t) = s(t) + n(t) \tag{4}$$

In formula (4), x(t) represents a noisy speech signal, S(t) represents a clean signal without noise, and n(t) represents a full noise signal in speech (Gao et al. 2012). We have improved the traditional repetitive Wiener filter and recycled linear prediction analysis to estimate $\hat{P}_s(\omega)_{i+1}$ for enhanced speech.

The improvement was modified as follows:

(1) Subtracting $\alpha\hat{P}_n(\omega)(a > 1)$ at a time frame with a higher spectral amplitude can better highlight the speech spectrum, suppress pure tone noise, and improve noise reduction performance. The current frame is taken out from the Wiener filtered signal, and the frame data is weighted by using the data of the previous frame of the silent segment to achieve dynamic updating of the noise spectrum of the silent segment.

(2) Since the distortion of the filtered spectrum is largely due to the difference between the noise power spectrum $P_n(\omega)$ and its estimated value $\hat{P}_n(\omega)$, the framed averaging of the noisy speech spectrum $X(\omega)$ can reduce the filtered distortion.

The Initialization improvements formula such as formula (5).

$$\begin{cases} \hat{P}_s(\omega)_0 = P_x(\omega) - \eta\hat{P}_n(\omega), P_x(\omega) - \eta\hat{P}_n(\omega) \geq \mu P_x(\omega) \\ \hat{P}_s(\omega)_0 = \mu P_x(\omega), P_x(\omega) - \eta\hat{P}_n(\omega) < \mu P_x(\omega) \end{cases} \tag{5}$$

The repetitive filter improvements formula such as formula (6).

$$\begin{cases} H(\omega)_i = \dfrac{\hat{P}_s(\omega)_i}{\hat{P}_s(\omega)_i + \hat{P}_n(\omega)}, \hat{P}_s(\omega)_i - \lambda\hat{P}_n(\omega) \geq \psi P_x(\omega) \\ H(\omega)_i = \dfrac{\psi P_x(\omega)}{\psi P_x(\omega) + \hat{P}_n(\omega)}, \hat{P}_s(\omega)_i - \lambda\hat{P}_n(\omega) < \psi P_x(\omega) \end{cases} \tag{6}$$

In our new model, we combine two complementary models: the optimized repeating Wiener filter model and Characteristics abandon model. During speech recognition, the new model will first use a repeated Wiener filter to filter the noisy speech. After the current frame data signal processed in weighting processing on a silent data. Accepting the disable MFCC speech feature, we use it as an input to the feature abandonment algorithm model. Finally, we use the feature abandonment algorithm model to recognize speech. After experimental verification, the sound noise has been greatly improved.

CONCLUSION

The object of this research is speech recognition in high noise in complex environments such as in automobile. We mainly improve and optimize the anti-noise and noise reduction capabilities of the system based on the ordinary speech recognition system. Through such improvement measures, the speech recognition system can be identified in a noisy environment. In the improvement, the combination of the discard feature method denoising and the enhanced speech method denoising in the anti-noise technology is adopted. This new method effectively eliminates the influence of noise on the speech recognition system, and greatly improves the recognition rate and accuracy of the speech recognition system under noisy environment.

ACKNOWLEDGEMENTS

Project Supported by Scientific and Technological Research Program of Chongqing Municipal Education Commission (No. kj1603810, kj1737456).

REFERENCES

Gao LY, Zhu W, Sang ZX, Mi L (2012) Speech enhancement algorithm based on improved spectral subtraction. Modern Electronics Technique, 35(17): 60-62.

Hu ZQ, Zeng YM, Zong Y, Li MC (2014) Improvement of MFCC parameters extraction in speaker recognition. Computer Engineering and Application. Computer Engineering and Applications, 50(7): 217-220.

Jiang JZ, Zhang DF, Zhang LH (2012) Speech enhancement algorithm for high noise environment. Computer Engineering and Applications. Computer Engineering & Applications, 49(20): 222-225.

Lee CC, Mower E, Busso C, Lee S, Narayanan S (2009) Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 53(9): 1162-1171.

Moataz EA, Mohamed SK, Fakhri K (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(2011): 572-587.

Wu SQ, Tiago HF, Chan WY (2009) Automatic speech emotion recognition using modulation spectral features. International Conference on Digital Signal, 53(5): 768-785.

Zhou YH, Li FL, Tong F (2013) The microphone array speech enhancement and HMM recognition joint processing in noise environment. NCMMSC' 2013, 47(6): 564-576.