The Software System Implementation of Speech Command Recognizer Under Intensive Background Nosie

IOP Conf. Series: Materials Science and Engineering **563** (2019) 052090 doi:10.1088/1757-899X/563/5/052090

The Software System Implementation of Speech Command Recognizer Under Intensive Background Nosie

SONGJingyu[1], CHENBo[1], JIANGKun[1], YANGMiaosheng[1] and XIAOXi[2]

[1]System Engineering Research Institute of China Shipbuilding Corporation, Beijing 100036, China

[2]Tsinghua University, Beijing 100084, China

Abstract. A device for speech command recognition is designed and implemented in this paper. The adaptive filtering technique is adopted for speech enhancement by using two microphones for sampling the background noise and noisy speech signal correspondently. The connected-word speech recognition algorithm based on HMM is employed for speech command recognition. The experiment shows that the equipment can work well under the background noise of 90dB.

1. Introduction

Speech command recognition technology for limited vocabulary based on conjunction recognition technology has good robustness and high recognition rate and it is widely used in many fields of information query and voice control.However, speech recognition technology is sensitive to environmental noise when it is applied, especially in strong noise environment, which has a significant impact on speech endpoint detection and feature extraction.Therefore, in a speech recognition system, speech enhancement technology is usually used to reduce the effect of noise on speech signal.The purpose of speech enhancement technology is to try to recover clean speech signals from noisy speech.Traditional speech enhancement methods mainly include parameter estimation methods such as comb filter (estimation of pitch period parameters), wiener filter, kalman filter (estimation of signal-to-noise ratio parameters) and parameter estimation-based speech enhancement methods such as spectral subtraction and adaptive filtering noise cancellation.In addition, there are voice enhancement methods based on microphone arrays and methods based on statistics and in-depth learning.Considering the size limitation of microphone array in application, the noise of application scene may be non-stationary, and the type and statistical characteristics of noise are unknown.In practical applications, we consider using non-parametric speech enhancement methods.

In non-parametric methods, spectral subtraction will produce serious spectral distortion under the condition of low signal-to-noise ratio, which will affect the performance of speech recognition. Therefore, we choose the adaptive filtering noise cancellation method of dual microphones as our speech enhancement method.In speech recognition model, we adopt HMM model based on segment length distribution, which has good adaptability and can achieve high recognition rate of command words under the condition of accurate speech endpoint detection. Finally, we transplant the speech enhancement algorithm and speech recognition algorithm to ADSP-21469 platform, and realize the speech recognition task in strong noise environment.

2. Implementation of speech enhancement and endpoint detection

Figure 1 is the block diagram for the realization of speech recognition algorithm.In the front of the speech recognition system, a speech enhancement scheme based on adaptive noise cancellation with dual microphones is used. The microphone labeled Mic_S in the system is close to the mouth to collect voice command signals containing environmental noise. Another microphone labeled Mic_N is set up to collect environmental noise.Under the condition of 90 dB ambient noise, the output signal-to-noise ratio of the microphone used to collect voice is about 0 dB under the condition of short talk.These two microphones acquire signals through adaptive filters to achieve noise cancellation function. The signal-to-noise ratio can be increased by more than 20 dB, which meets the requirement of speech signal-to-noise ratio of follow-up command word speech engine.The speech recognition module of command words used in the system includes endpoint detection, speech feature extraction and template search matching. The final recognition result is sent to the controller system through the data communication interface. Here we introduce the functions and implementation algorithms of the main modules of the system in detail.

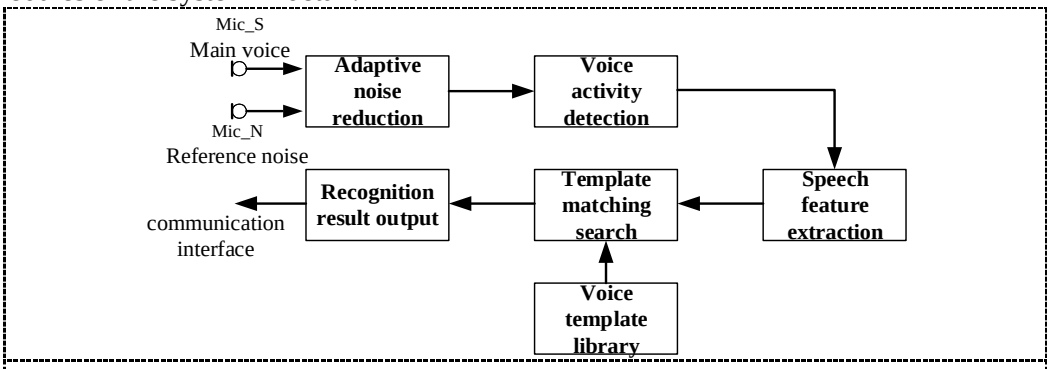

Figure 1. Block diagram of voice recognition.

2.1. Speech enhancement algorithm based on adaptive filtering

Figure 2 isthe block diagram of adaptive filter in speech enhancement principle used in the system. Mic_N is used to collect background noise signal $N(n)$ and Its output $x(n)$ is input to the adaptive filter as a noise reference source.The signals collected by microphone Mic_S include ambient noise $N'(n)$ and speech signals $S(n)$.In practical applications, in order to cooperate with the adaptive filter algorithm, it is necessary to define a delayer τ to compensate for the effects of $N(n)$ and $N'(n)$ noise path difference and adaptive filter delay in microphone output.The purpose of this method is to make the adaptive filtering noise cancellation algorithm better adapt to various noise environments.

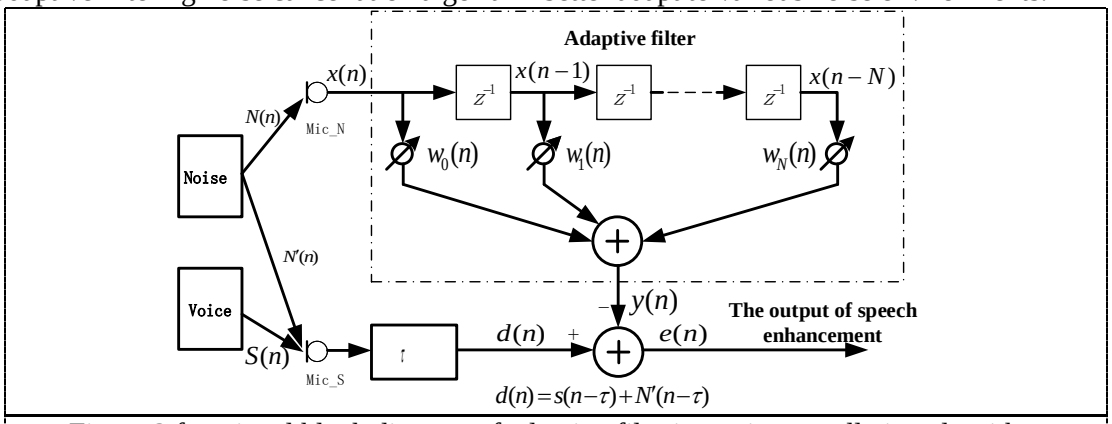

Figure 2.functional block diagram of adaptive filtering noisecancellation algorithm

In Figure 2, the adaptive filter is in transverse form.Its input signal is $x(n)$. The input vector $x(n)=\left[x(n),x(n-1),\cdots,x(n-N)\right]^T$. The weighted coefficients of adaptive filters are $w_i(n)$, $i=0,\cdots,N$. Record in vector form, $W(n)=\left[w_0(n),w_1(n),\cdots,w_N(n)\right]^T$. Output of adaptive filteriscalculatedas follows.

$$y(n)=\sum_{i=0}^{N} x(n-i)*w_i(n)=W^T(n)\bullet X(n) \tag{1}$$

$d(n)$ is the output signal of the delayer, $d(n)=s(n-\tau)+N'(n-\tau)$, containing the original speech signal and environmental noise signal. Error signal by subtracting it from the output of the adaptive filteris the enhanced speech signal, $e(n)=d(n)-y(n)$.

The purpose of the adaptive filter is to enable the output $y(n)$ of the filter to accurately cancel out the environmental noise component $N'(n-\tau)$ contained in $d(n)$. Using the least mean square error estimation algorithm, it is necessary to estimate the weighting coefficient $w_i(n)$ of the filter in real time so as to minimize the expected value $E\left\{e^2(n)\right\}$ of the error. In this case, both environmental noise $N(n)$ and $N'(n)$ are homologous noise, and they are almost statistically unrelated to speech signal $s(n)$.It satisfies the following conditions.

$$E\left\{N(n)s(n)\right\}\approx 0,\quad E\left\{N'(n)s(n)\right\}\approx 0 \tag{2}$$

In this system, the Mic_S microphone for voice acquisition is close to the speaker's mouth, and the Mic_N for background noise acquisition is relatively far from the speaker's mouth.Because of the near-field effect of speech, the signal-to-noise ratio of speech collected by Mic_S microphone is much larger than that collected by Mic_N microphone. That is to say $x(n)\approx N(n)$, the signal collected by Mic_N microphone in strong noise environment can be approximately considered as environmental noise.In this case, the condition of formula (3) is approximately satisfied.

$$E\left\{x(n-i)s(n-\tau)\right\}\approx 0,\quad i=0,\cdots,N \tag{3}$$

According to formula (3), when estimating the minimum mean square value, the following is derived.

$$E\left\{e^2(n)\right\}=E\left\{\left[s(n-\tau)+N'(n-\tau)-y(n)\right]^2\right\}\approx E\left\{\left[s(n-\tau)\right]^2\right\}+E\left\{\left[N'(n-\tau)-y(n)\right]^2\right\}$$

$$E\left\{e^2(n)\right\}=E\left\{\left[s(n-\tau)+N'(n-\tau)-y(n)\right]^2\right\}$$

$$\approx E\left\{\left[s(n-\tau)\right]^2\right\}+E\left\{\left[N'(n-\tau)-y(n)\right]^2\right\} \tag{4}$$

When $E\left\{e^2(n)\right\}$ reaches the minimum value, it shows that the noise $N'(n)$ in noisy speech is accurately cancelled by the output $y(n)$ of the filter, and then the output $e(n)\approx s(n-\tau)$ of speech enhancement approximately recovers the clean speech.

We use the steepest descent method to find the tap coefficients that minimize the mean square error, which minimizes the error function $\xi = E\left\{e^2(n)\right\}$. For LMS algorithm, in practical application, we can not get the set average information of errors, so we use instantaneous error $\xi = e^2(n)$ instead of estimating the set average of errors.Similar to the steepest descent method, we get the following formula.

$$W(n+1) = W(n) - \mu(n) \bullet \nabla\left[e^2(n)\right] \tag{5}$$

In formula (5), $\mu(n)$ is the step-size coefficient of the algorithm, and gradient operator ∇ is concerned with finding partial derivatives of W . Its vector form is expressed as follows.

$$\nabla = \left[\frac{\partial}{\partial w_0(n)}, \frac{\partial}{\partial w_1(n)}, \cdots, \frac{\partial}{\partial w_{N-1}(n)}\right] \tag{6}$$

Notice that $e(n) = d(n) - y(n)$ and $d(n)$ is independent of W , so $\frac{\partial\left[e^2(n)\right]}{\partial w_i(n)} = -2e(n)\frac{\partial y(n)}{\partial w_i(n)} = -2e(n)x(n-i)$ and formula (5) can be written as follows.

$$W(n+1) = W(n) - 2 \bullet \mu(n) \bullet e(n) \bullet X(n) \tag{7}$$

This is the iterative formula of LMS algorithm. The convergence and convergence speed of LMS algorithm are controlled by its iteration step coefficient μ. Only when μ is small enough can LMS algorithm guarantee the convergence stability of this iteration algorithm. When the iteration is in a stable convergence state, the mean error gradient is equal to 0, which corresponds to the optimal filter coefficient W_0 and the minimum filter output error e_{min}^2 . In the real-time iteration process, the instantaneous error gradient is used to update and estimate the filter coefficient $W(n)$. Therefore $W(n)$ is perturbed around the optimum coefficient W_0, the error $e^2(n)$ of the filter has an additional error e_e^2 besides the minimum error e_{min}^2 . That is to say $e^2(n) = e_{min}^2 + e_e^2$. The offset coefficient $M = \frac{e_{excess}^2}{e_{min}^2}$ is defined in the LMS algorithm to evaluate the convergence state of the adaptive filter. When the value of M is small, for example $M < 0.1$, the misalignment coefficient has the following approximate relationship.

$$M \approx \mu \bullet tr\left[R\right] \tag{8}$$

It can be seen from formula (8) that the larger the step size coefficient μ , the larger the misalignment coefficient, which will eventually lead to greater output error of the filter, leading to instability.The analysis of filter[1] shows that if signal $x(n)$ is stationary, then the autocorrelation matrix of signal $x(n)$ is $R = E\left[X(n)X^T(n)\right]$, and then $0 < \mu < \frac{1}{3tr\left[R\right]}$ is a sufficient condition for convergence of LMS iteration algorithm.Here the trace of the autocorrelation matrix R is $tr\left[R\right] = \sum_{i=0}^{N} x^2(n-i)$.It can be seen that in LMS algorithm, the value of iteration step coefficient

$\mu(n)$ is related to the magnitude of the input signal and needs to be adjusted dynamically. If $\mu(n)$ is set as a constant, it will affect the stability and convergence speed of LMS algorithm in practical application.To solve this problem, many improved algorithms have been proposed. We use the improved normalized LMS algorithm (NLMS).The NLMS algorithm takes into account the change of input signal amplitude and normalizes the calculation of iteration step coefficients.Its iteration formula is as follows.

$$\mu(n) = \frac{1}{2 \bullet X^T(n) \bullet X(n)} \tag{9}$$

$$W(n+1) = W(n) + \frac{1}{X^T(n) \bullet X(n)} e(n) \bullet X(n) \tag{10}$$

NLMS algorithm can be seen as optimizing u again on the basis of formula (7), so that the posterior error of filter coefficients calculated from formula (7) is minimized when applied to transverse filters[1].

$$\hat{e}(n) = d(n) - W(n+1)^T \bullet X(n) \tag{11}$$

Experiments show that the NLMS algorithm takes into account the stability and convergence speed of adaptive filtering, and the algorithm is relatively simple and the computational complexity is small.

Speech endpoint detection algorithm

After noise cancellation, the signal-to-noise ratio (SNR) of speech signal is enhanced, which can satisfy the basic requirement of endpoint detection for SNR, so that the position of speech signal can be detected more accurately from background noise.

Experiments show that the speech recognizer is sensitive to the initial position error of speech. The software of the recognizer uses a double threshold detection method based on short-term average energy to determine the starting point of speech more accurately.Since voice always has a high energy voiced sound after the beginning of speech, a higher threshold T_h is set to judge the location of voice N_0. In order to avoid the loss of data with less energy at the beginning of voice segment, a threshold T_i lower than T_h is set to detect the real voice starting position N_1.As shown in Figure 3, the double threshold detection method is based on the correct estimation of background noise energy. In endpoint detection algorithm, cyclic buffer is used to cache voice data.In order to estimate the background noise of speech signal after speech enhancement, the algorithm calculates the energy of input signal of one frame (frame length is 20ms) every 10ms interval, and caches the speech data in a circular queue of 500 frames (5 seconds). In speech detection, the minimum energy of all data frames in the cyclic queue is searched as the background noise energy T_{noise} of the current speech. Setting two thresholds $T_h = T_{noise} + 20dB$ and $T_i = T_{noise} + 6dB$.When detecting the location N_0 where the frame energy of speech is higher than the threshold value of T_h, it is considered that a speech event is detected, and the frame energy value of speech is traced back from that time until the location where the frame energy is less than the threshold of T_i is found, which is the real starting time of speech N_1.In addition, in order to prevent misjudgement in retrospective speech starting point, it is agreed that the maximum number of retrospective frames should not exceed 20 frames. When the N_0 position is detected, the voice pronunciation begins. The frame energy of the voice lasts more than 0.5 seconds and falls again below the threshold value T_i. This is the end of the voice pronunciation, that is the N_2 position in Figure 3.The method of speech detection based on energy threshold alone works well in quiet environment. However, when there is background noise, this method is prone to interference, resulting in speech error judgment. For this reason, the analysis of pitch features is added in this system. Its principle is that there are voiced components in the pronunciation of Chinese

syllables.A remarkable feature of voiced tone is that there is a relatively stable fundamental frequency, so we add a pitch search algorithm to the algorithm.We search for pitches in the range of 3-13 ms pitch period. If more than five pitch frames are found in succession, the voice will be recognized as voice, otherwise it will be considered as interference signal. Using pitch features to detect speech greatly improves the stability of the endpoint detection algorithm.

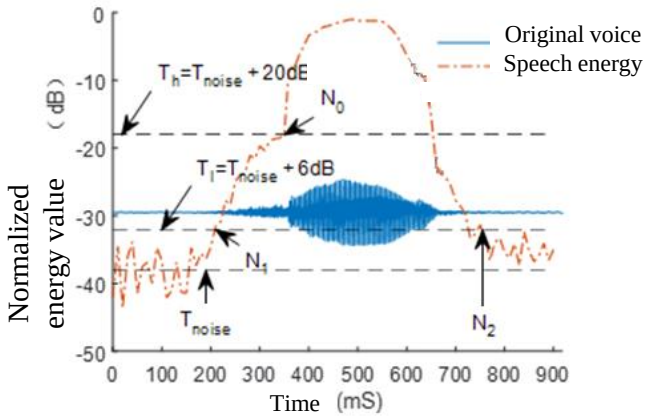

Figure 3. Short-term Energy Double Threshold Voice Power-off Judgment Diagram.

3. Command word recognition algorithms and implementation scheme

In the recognition system, the sampling frequency of speech signal is 16 KHz, the adaptive filter uses 320-point FIR transverse filter, and the speech microphone channel inserts 160-point delay to automatically adapt to the change of the two-channel noise delay path environment. After speech enhancement, the speech endpoint is judged, and then the speech recognition features are extracted.Speech signal is pre-emphasized with a high-pass filter $H(z)=1-0.98Z^{-1}$. The speech signal is analyzed by using Hamming window. The frame length is 20 ms and the frame is shifted to 10 ms. The speech feature vector is a 45-dimensional feature vector composed of normalized energy and 14-dimensional MFCC coefficients, together with their first-order and second-order difference coefficients. The system uses DDBHMM [2] model as the basic recognition framework, and uses two Gauss functions of total covariance matrix to form GMM model to describe the distribution of observation vectors.Considering the flexibility and stability, the system uses half-syllable (each Chinese syllable can be divided into Initial and Final parts) as the modeling primitive for word stitching. The system uses 100 Initial models and 164 Final models. Each Initial model is described by two states, and each Final model is described by four states. There are 856 states in the whole model. The training of the HMM model used the soliton full syllable database (1254 syllables) of 100 people (50 male voices + 50 female voices) and the connection word database (699 words) of 100 people (50 male voices + 50 female voices). The two databases covered all the connections between the commonly used syllables and syllables in Chinese.HMM model training is carried out on PC in advance, and the trained model is stored in Flash chip through the simulator development interface.

When recognizing DDBHMM model, the frame synchronization search algorithm based on Viterbi decoding can be used directly to recognize the DDBHMM [2]model without considering the correlation of segment length between states, the correlation between observation vectors and the uniform distribution of segment length. If the recognition algorithm is searched directly according to the MLSS recognition algorithm of DDBHMM model, the calculation of the Gauss probability model of the total covariance matrix is very large. Therefore, a two-stage search strategy is adopted.

The first stage is the fast search stage.The state codebook of HMM uses a single Gauss diagonal matrix to find the top N candidate entries with the highest matching degree.The second stage is precise search stage, using GMM model of total covariance matrix.The likelihood distance is calculated

according to MLSS recognition algorithm for the N candidate entries which are searched in the first stage, and the best candidate entries are obtained.

The hardware system of the recognizer uses ADSP-21469 DSP [3] chip to form the core of hardware processing. This is a 32-bit floating-point DSP, which is very convenient for the transplantation of speech recognition algorithm.The hardware design of the system refers to the ADSP-21469 EZ-KIT Lite [4] evaluation board provided by AD Company, and cuts it. Flash with 32M bytes and memory system with 128M DDR2 bytes are designed. The system board also designs a serial port for data communication.Considering that the system needs to work in a strong noise environment and the voice acquisition system needs a large dynamic range, a 24-bit ADAU1361 Codec is designed as the analog front-end circuit of voice acquisition on the DSP board.

The software of this system is debugged and compiled by AD's Visual DSP C++ 5.0[5] compiler system and downloaded to a special DSP processing board to run without operating system support.

4. The recognition test results of command word system

The speaker is used to simulate the noise source in the test. The system microphone Mic_N is located at the top of the helmet to collect background reference noise. It is close to the noise source (speaker). At this time, the maximum sound pressure level at the microphone Mic_N can reach more than 100 dB.The microphone Mic_S speaks close to the mouth to collect voice. In the test, Mic_N and Mic_S microphones are more than 20 cm apart. In the experiment, we first test the effect of speech enhancement system, and then test the speech recognition.

4.1. Speech enhancement algorithms effect test

We tested the effects of single-tone positive selection, dual-tone multi-frequency signal and aero-engine noise on speech enhancement.

4.2. Comprehensive testing of command word recognition system

The command words used in the test are all four-character words, totaling 100 command words. The recognition time of the DSP system is less than 0.5 seconds. Table 1 is the test results of the DSP hardware system in various noise environments.

Table 1.Real-time recognition test results of DSP hardware system.

Noise environment	*Recognition rate*
Quiet Office	95%
70 dB Noise of Aircraft Engine	92%
440Hz Single Frequency Pure Tone Noise 90dB	91%
Aircraft Engine Noise 90dB	90% (Refusal to recognize)
440Hz Single Frequency Pure Tone Noise 100dB	90%
Aircraft Engine Noise 100dB	85% (Refusal to recognize clearly)
Aircraft Engine Noise 110dB	Speech Detection Difficulties
440Hz Single Frequency Pure Sound Noise 110dB	85%

5. Conclusion

This DSP speech recognition system can perform normal speech recognition in quiet environment and engine noise environment with noise less than 70 dB. The recognition rate is over 90%.In the noise environment of engines larger than 90 dB, it is difficult to detect the endpoint of speech, which is manifested by refusing to judge the speech.When the noise is greater than 100 dB, the system refuses to recognize more clearly, and can not work normally at 110 dB.

Experiments show that speech recognition in strong noise environment is still a very challenging research topic. In the engine noise environment, speech enhancement and recognition research still has many aspects that can be improved.

References

[1] B Farhang-Boroujeny 2013 *Adaptive filters: theory and applications* (New York: Wiley–Interscience)
[2] Wang Zuoying and Xiao Xi 2004 *Journal of Electronic Science*. 32(1)
[3] Ji, Peifeng, W. Hu, and J. Yang 2016 *Ultrasonics.* **67**160-167.
[4] Horowitz Gary T and Hubeny Veronika E 2000*Journal of High Energy Physics*. (2000):031.
[5] TangJinshan 2004 *Digital Signal Processing.* **14.3**(2004):218-226