Speech noise Reduction and Automatic Recognition Based on Active Detection Algorithm

Jianli Mo¹, Chenyu Wu²

¹Electronic Information Engineering, Hunan University of Information Technology, ChangSha, HuNan, 410151, China
²Internet of Things, Hunan University of Information Technology, ChangSha, HuNan, 410151, China

*Corresponding author’s e-mail: 924392998@qq.com

Abstract: In practice, people need to carry out intelligent speech analysis on the sound collected or recorded in some specific scenes. These sounds are recordings of a period of time in some specific places. Due to the influence of the phonetic environment and the limitation of acquisition or recording means, the actual effective and available voice content is very small, and the voice quality is relatively poor. Therefore, it is necessary to enhance the speech of the acquired recordings, eliminate the noise and the interference of non-human voice, and improve the understanding degree of the target's voice. Voice quality is improved, and the voice can be targeted content recognition and keywords detected, the audio files into a text file, by setting the key words found early warning, the valuable data information in mass speech quick preview, retrieval, analysis, according to the established goals and objectives by using automatic speech recognition in order to improve work efficiency.

1. Introduction
The voice noise reduction based on active detection algorithm and automatic identification technology, can break through the non-stationary noise environment single channel speech enhancement methods the current situation of poor performance, to improve the noise reduction performance, using keywords to identify and continuous speech recognition technology at the same time, from a large number of find sensitive to conform to the goal of business speech, speech to improve speech recognition accuracy. The main functions and technical indexes of speech noise reduction and recognition can achieve: (1) speech noise reduction and enhancement. For the speech enhancement effect in noisy environment, the SNR can be improved by about 10dB; (2) speech detection. For the actual acquired signals, the accuracy of speech detection can reach 90%; (3) keyword detection. For speech detection and noise reduction enhancement, the detection accuracy of Chinese keywords is no less than 85%; (4) in terms of continuous speech recognition, the accuracy rate of Chinese transliteration is no less than 75% for basically audible speech after speech detection and noise reduction enhancement.

2. Speech Enhancement Technology
The speech enhancement technology adopts the advanced voice activity detection algorithm, noise estimation algorithm and noise suppression algorithm, of which the voice activity detection algorithm can be used to accurately detect the initial position of the target voice in a complex environment. It can
provide important references to the subsequent noise estimation. The noise estimation algorithm can dynamically track the changes of non-steady noise of different frequency bands within the scope of broadband and thus has very high accuracy in noise estimation. The noise suppression algorithm can effectively reduce background noise, highlight voice elements in the signal and improve the understandability of the voice.

Speech enhancement is mainly consisted of four parts, i.e. voice activity detection, noise estimation, noise suppression and post-processing. The specific contents are as follows:

1) Voice activity detection technique: In order to solve the common problem of detection inaccuracy of target voice in a strong noise environment, we can adopt multi-feature fusion methods that include energy features, zero-crossing rates and spectral correlation characteristics to estimate the probability of voice existence using probability model based on the Bayesian theory. On this basis, we can use the data-driven approach to get the optimum threshold to output the voice activity detection result or directly output the probability of voice existence as the soft-decision result of voice activity detection.

2) Noise estimation technique: By applying the improved minimum constraint recursive average algorithm and the spectrum estimation technique, we can accurately obtain background noise in dynamic change. We firstly use the above-mentioned voice activity detection method to estimate the probability of voice existence. Then we use the spectrum smoothing technique to remove strong voice elements in the signal, which will effectively improve the accuracy of noise estimation. Finally, we get dynamic background noise estimation according to the non-continuous feature of the voice signal.

3) Noise suppression technique: The system uses the non-causal prior SNR estimation method to enable noise suppression of the voice and estimate the probability of voice existence in different time periods and of different frequency bands. It further uses the estimation result to suppress noise so as to ensure the quality of voice segments and voice bands and maximally remove the noise from non-voice segments and non-voice bands. The non-causal feature of the system also ensures its real-time data processing.

4) Post-processing technique: We amend and supplement the result of noise suppression algorithm to clarify the voice note and strengthen the voice signal.

3. Keyword Recognition

The retrieval algorithm is the core of the keyword retrieval technique. It adopts the keyword retrieval strategy that is based on confusion network. The algorithm framework is shown as follows:

![Figure 1. An overall framework of keyword retrieval system](image)
In the picture above, the whole retrieval system is composed of three stages and the specific contents are as follows:

3.1. Acoustic model modeling technique
Among acoustic model modeling techniques, MPE is currently a most efficient identification training method of acoustic model. It has better performance than traditional training methods that is based on maximum likelihood estimation. On the basis of standard MPE training algorithm, we will adopt the following algorithm to improve it with the features of spontaneous speech.

In the comparison of phoneme accuracy, standard MPE takes the middle phoneme of the logical 3-factor model as the baseline for comparison. However, since real acoustic model is a clustering model that is measured in states, the tradition comparison method tends to give higher phoneme accuracy than it actually is, so it’s inaccurate. Improved MPE-SC algorithm takes the corresponding physical state sequence of 3-factor model as the baseline for comparison of accuracy calculation, thus it gives more accurate phoneme accuracy.

MPE-PPS algorithm introduces smoothing factor to the posterior probability calculation in standard MPE training to improve the condition where posterior probability is overly concentrated in distribution between two extremes 0 and 1, thus improving confusion among competitive factors and effectively increasing MPE’s update statistical magnitude. Larger statistical magnitude can provide more robust model estimation so as to improve model performance.

3.2. Decoding algorithm
In terms of decoding algorithm, we can adopt the approach as follows:

We will use the Cross-word static search spatial construction method that is based on state network to statically compile the lexical tree, phoneme context and acoustic model into search space to form a memory-saving state network. In terms of linguistic model, we will enable dynamic extension by introducing two layers of hidden nodes to cluster in cross-word context by fanning-in and fanning-out triphone. We can use the layer of WI to store the identification number of words and put this layer of nodes in the middle part of non-shared network. In the merging process of forward/backward nodes, the network state is optimized at the state level, and the network is organized into a multilayered structure to ensure efficient token propagation in every single layer.

3.3. Confusion network-based keyword retrieval algorithm
The core of keyword retrieval algorithm is the generating and searching of confusion network. A standard voice recognition confusion network can be represented as the vector of confusion class \( S_N = \{s_1, \ldots, s_n, \ldots, s_N \} \), of which \( N \) is the number of confusion class in confusion network. As for a keyword \( W_M \), by looking up the dictionary, we can get its corresponding syllable sequence \( Q_M = \{q_1, \ldots, q_m, \ldots, q_M\} \), of which \( M \) is the number of syllables of the keyword \( W_M \).

The basic concept of forward search algorithm, as is shown in the picture, is matching the syllable sequence of every keyword \( Q \) with the slicing of confusion network \( S \), from the first slicing to the last slicing, to get the maximum matching probability using dynamic programming algorithm. Then make comparisons with all keywords. The corresponding keyword of the maximum matching probability will be optimal detection candidate, which is as following:

\[
Q_{\text{best}} = \max_{Q_M} R(Q_M, N)
\]

(1)

There are many NULL sides in the generation process of confusion network. Thus the process of keyword searching is a classical dynamic programming problem. Assume that at the confusion class
$s_n$, the syllable sequence of optimal matching keyword is $q_1d_1q_2d_2\ldots q_md_md_M$, of which $d_m$ is the number of NULL sides before $q_m$. Assume $R(Q_M,n)$ is the average matching score of $Q_M$, then define it as follows:

$$R(Q_M,n) = \frac{1}{M} G(m,n)_{|m=M}$$

(2)

(Here, $G(m,n)$ is the cumulative score of $Q_M$ at $q_m$, and the calculation formula is as follows:

$$G(m,n) = \max_{d_m} \{G(m-1,n-d_m-1)+C(m,n)+\text{penalty}(d_m,n)\}$$

(3)

Here, $\text{penalty}(d_m,n)$ is penalty factor of syllable $q_m$:

$$\text{penalty}(d_m,n) = \ln \left\{ \prod_{k=1}^{d_m} p(\text{null} \mid s_{n-d_m-1+k} ) \right\}$$

(4)

$C(m,n)$ is the matching score of $q_m$ and $s_n$:

$$C(m,n) = \log \{ (1-\alpha) p(q_m \mid s_n, O) + \alpha P_{\text{conf}}(m,n) \}$$

(5)

$$P_{\text{conf}}(m,n) = \sum_{\{q \in \text{FuzzySet}(q_m)\}} p(\varphi_i \mid s_n, O) p(q_m \mid \varphi_i)$$

(6)

Here, $p(q_m \mid s_n, O)$ is the posterior probability score and $\alpha$ is the regulatory factor. $P_{\text{conf}}(m,n)$ is the confusion degree of $q_m$ and $s_n$ calculated according to the confusion matrix, and $p(q_m \mid \varphi_i)$ is the syllable similarity of $q_m$ and $\varphi_i$, and $\text{FuzzySet}(q_m)$ is the syllable set that is hard to tell apart from $q_m$.

Figure 2. confusion network –based keyword retrieval algorithm
4. Continuous Voice Recognition
As for the massive voice in complex environments like network space, voice recognition technique can be used to quickly recognize and identify the content of voice. The most used technique is the acoustic and linguistic modeling technique of voice data that can effectively identify multifarious network language for large populations. Its core technologies mainly include:

4.1. Acoustic modeling that is based on Deep Neural Network (DNN)
In the framework of conventional voice recognition, the Gaussian mixture model (GMM) is often used in the statistics and modeling of the probability distribution of acoustic features.

As for small-scale corpus, GMM model can fit sounding phenomena in different phoneme states quite well. Thus, it’s hard for the conventional GMM model to accurately describe and fit the probability distribution of acoustic features, and the accuracy of voice recognition dropped rapidly. To solve this problem, the Deep Neural Network structure can be applied to the modeling of the distribution of acoustic features. Thanks to the powerful nonlinear fitting ability of the Deep Neural Network, the DNN acoustic model can accommodate all kinds of acoustic probability distribution. In the training process, we can also use the GMM-HMM model to ensure voice frame alignment of all voice date in advance so as to get the status tab of voice frame data. In terms of training lose function, the minimized state classification error can be taken as the training criterion which can not only quicken the model’s rate of convergence, but also improve its classification ability and voice recognition accuracy.

4.2. The choice of colloquial language model based on the degree of confusion
Because of VOIP phenomenon, there are many colloquial words in the language vocabulary distribution, and written language has great difference, for voice and data suitable for network language model, can harvest a lot of Internet spoken text data, and the “N –Gram” algorithm is adopted to the vocabulary of the text data statistical modeling, in the process of training, the parameters of the model is obtained by using the maximum likelihood estimate. In training, different N values will correspond to different language models. The larger N value is, the more complex the language model will be and the decoding efficiency will be reduced. In order to select the best language model, this paper USES the degree of confusion to measure the performance of different language models, and sorts and screens them according to the degree of confusion of all candidate models. The smaller the degree of confusion, the better the effect of the language model.

4.3. Decoder optimization based on weighted finite state machine(WFST)
In the process of speech recognition, a static decoding network can be constructed according to the existing acoustic model, phoneme context, dictionary and language model, which is used to calculate and query the probability of decoding path. Due to the large size of the colloquial language model, constructing the static decoding network directly using the optimal language model will occupy a large amount of memory space, and lead to a large number of candidate edges or paths in the decoding
process of Token Passing algorithm, which seriously affects the time efficiency of the decoder. A simple language model is used to construct the basic static decoding network, and the token is passed based on the decoding network. The intermediate result of WFST token passing is represented as Lattice, and finally complex language model is used to conduct re-scoring operation on the path of lexical Lattice. By combining simple and complex language model and decoding together, the speed of decoding can be greatly improved without obvious loss of recognition accuracy.

4.4. File management application system
File compression and storage: apply appropriate digital signal processing technology to the original digital audio signal stream, reduce (compress) its bit rate for encrypted storage under the condition of no loss of useful information or negligible loss introduced. After intelligent analysis, the speech files were labeled and classified, and the newly generated speech files were mapped to the original files.

5. Conclusion
In this paper, aiming to check the target content of the optimized speech is identified and the key words, the speech files are converted into text files, and the valuable information in mass speech is quickly previewed, retrieved and analyzed by setting the keyword warning, so as to improve the work efficiency and reduce the complex work intensity. This project uses digital signal processing technology, speech noise reduction and enhancement technology, speech detection technology, and speech automatic recognition technology to carry out noise reduction and enhancement processing of acquired speech information, extract pure original speech signals, improve the clarity of speech, improve the SNR, and make speech easier to understand and recognize.

Acknowledgements
This work was supported in part by some world's leading intelligent science and technology research institute, the Fundamental Research Funds for the Central Universities with Grant N172304022, the Science and Technology Plan Project of Guangzhou with Grant 201804010433, and the Bidding Project of Laboratory of Language Engineering and Computing with Grant LEC2017ZBKT001.

References
[1] Chen, F., Gao, S. (2010) Speech recognition technology and development. Telecommunications Science, 10, pp.54-57.
[2] Lei, J. (2002) Research and implementation of speech recognition technology. Wuhan University of Technology.
[3] Bao, P.X. (2016) Research on phase noise suppression algorithm in SCMA-OFDM system. Southwest Jiaotong University.
[4] Ma, Z.Q., Li, T.Y., Yang, S.T., et al. Mongolian acoustic model modeling based on deep neural network. Journal of Intelligent Systems, vol.13, no.71, pp.160-166.
[5] YOUNG, S. (1994). Tree-based state tying for high accuracy acoustic modeling[C]/ Proc. ARPA Human Language Technology Workshop.
[6] BONEH, D. (2004) Public key encryption with keyword search.
[7] Hristidis, Vagelis, Hwang, Heasoo, Papakonstantinou, Yannis. Authority-based keyword search in databases. Acm Transactions on Database Systems, vol.33, no.1, pp.1-40.
[8] J.H. Conway, & N.J.A. Sloane. (1982). Fast quantizing and decoding algorithm for lattice quantizers and codes. Information Theory IEEE Transactions on, vol.28, no.2, pp.227-232.