http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 8 (5): 796-800, 2009
ISSN 1812-5638

Speech Recognition Algorithm of Parallel Subband HMM Based on Wavelet Analysis and Neural Network

Zhou Ping, Tang Li-Zhen and Xu Dong-Feng
Guilin University of Electronic Technology, Guilin 541004, Guangxi, China

Abstract: The purpose of speech recognition is able to extract the content of the speech in various conditions and transform from speech to text automatically. Based on human hearing perception mechanism, we propose a decomposition method using independent parallel subbands for speech recognition. In this method, wavelet processing is introduced into Hidden Markov Model (HMM) and Fuzzy Neural Network (FNN) is used to improve the convergence speed and to avoid local optimal in speech detection. Experiment results show that the proposed hybrid speech recognition model is more robust when noise presents. We combine the dynamic modeling of CDHMM and the classification capability of fuzzy neutral network, this has been a hot research area in recent years and can be applied to speech-text transform devices.

Key words: CDHMM, wavelet, independent parallel subband, fuzzy neural network

INTRODUCTION

The purpose of speech recognition is to make machines understand human languages, which means they are able to extract the content of the speech in various conditions and transform from speech to text automatically (Douglas, 2003). In recent years, Continuous Density Hidden Markov Model (CDHMM) has been used as the major technology for speech recognition (Xinwei *et al*., 2005). It is a dynamic information processing method based on cumulative probabilities of timing sequence and has the advantage of modeling dynamic time sequence (Tsung-Hui *et al*., 2008). When used together with wavelet (Deshpande and Holambe, 2008), CDHMM has a better performance in the differentiating between speech signals and background noises.

When Fletcher did his research on the human auditory perception of telephone speech, he found that human hearing system has product rule of misrecognition ratio on narrow signals i.e., the misrecognition ratio of bandpass speech signal is the product of misrecognition ratio on different frequency bands. Based on the previous research results, Allen has summarized that perception of phoneme (Bershtein and Kovalev, 2004) in hearing is the combination of information which has been processed individually on different frequencies (Allen, 1994). This is known as Fletcher-Allen algorithm. In this study, we propose a speech recognition method which makes use of independent subbands on top of Fletcher-Allen algorithm.

During the last decade of the 20th century, the neural network has become a new method for speech recognition (Desai *et al*., 2003) owe to its advantages of self-organizing, self-adaptation and self-learning. The fuzzy neural network (Bershtein and Kovalev, 2004) proposed in this study may avoid local optimal and is suitable for speech recognition, which is a simulation process of human intelligence and is hard to model in algorithm. However, its performance in time sequence is less than satisfactory. By combining the dynamic modeling of CDHMM and the classification capability of fuzzy neural network, this allows synergistic effects of both models in speech recognition. There has been an increasing trend in such combination in this area of research (Trentin and Gori, 2003).

PARALLEL SUBBAND HMM AND NEURAL NETWORK ALGORITHM ON WAVELET ANALYSIS

Performance analysis of subband algorithm: Fletcher-Allen algorithm shows that the misrecognition ratio on an entire band is the product of misrecognition ratios in all subbands. Therefore, we aim to establish individual recognition model for the different subbands, the results of each recognition model will then be merged into the final output of recognition result.

In some cases, even there are poor detections at certain frequencies, a system could still get a very good overall detection provided that there are good detections at the other frequencies.

Corresponding Author: Zhou Ping, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China

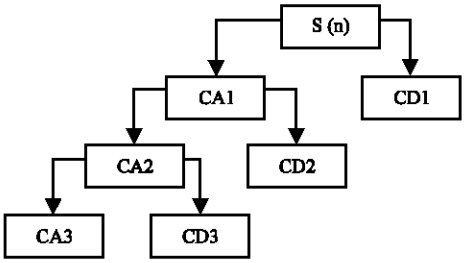

Fig. 1: The three-stage decompose of the signal

Terminal detection and noise suppressing using wavelet: Wavelet processing is widely used for noise elimination, in order to improve the terminal detection and speech recognition quality (Sungwook *et al*., 2001). Quadrature wavelet transform makes use of fast wavelet transform, which can remove the correlation among signals and concentrate signal energy. Therefore, through wavelet transform, signal power can be converted to the coefficients of several certain sub-frequencies. When the coefficients of other sub-frequencies are set to 0 or small weights, noise can be suppressed efficiently. The noise can be suppressed in the following procedures: first decompose the signal in three-stage (Fig. 1 shows the process of decomposition), since noises are contained in the high frequency of CD1, CD2 and CD3, we may then process the coefficient using a threshold value to remove noises and to get the real speech signal s(n).

After removing the noise by wavelet transform, we can then calculate the instantaneous energy E_n and average zero crossing ratio Z_n. The start and end points of a speech can be found when measuring the terminal using VUS algorithm.

$$E_n = \sum_{k=0}^{W_L-1} s_n^2(k) \quad (1)$$

$$Z_n = \frac{1}{2}\left[\sum_{k=0}^{W_L} |\mathrm{sgn}(s_n(k)) - \mathrm{sgn}(s_n(k-1))|\right] \quad (2)$$

A new neural network: The artificial neural network has the features of self-adaptive, self-organizing and self-learning. However, it has two major drawbacks. First, the training speed is slow and it sometimes converges on the local optimal points of the target function. Secondly, the network is analogous to a black box in which the weights might not have clear meanings, it cannot make use of previous knowledge to improve the network structure and to speed up learning process by avoiding local optimum points. If we merge Fuzzy Logic (FL) with various types of neural network (Leung *et al*., 2003), which is known as the Fuzzy Neutral Network (FNN), then both can make use of their advantages (Pulasinghe *et al*., 2004).

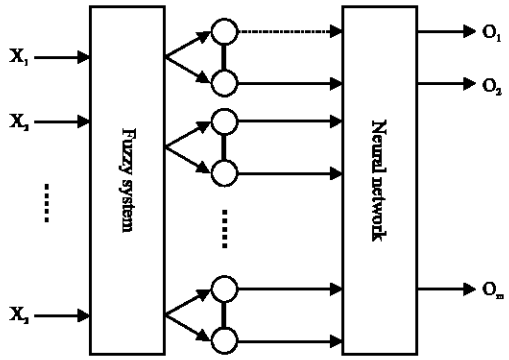

Fig. 2: Fuzzy neural network system

The fuzzy neural network put forward the concept of membership function, which could map an element to a proper degree of membership. This is ideal for speech recognition which has to deal with objects with no distinct boundaries, i.e., fuzzy relations (Lilia and Sellmai, 2003). Therefore, when we use membership function to describe the fuzzy relations, performance of the recognition system would be improved (Xiying and Guozhong, 2007).

As noises are present in the environment, direct use of the eigenvectors of the input signal will produce poor detection results. If we first apply fuzzy operator to the eigenvectors followed by fusion and classification, performance of the recognition can be improved significantly. Figure 2 shows an example of a series FNN. It has n input values and m outputs. The original inputs, i.e., $\{x_i, \mid i = 1, 2, L, n\}$, will be used as inputs of the neutral network after fuzzing, which has a total number of inputs of n*l. The output of the system is $\{o_k, \mid k = 1, 2, L, m\}$.

PARALLEL SUBBANDS CDHMM AND FUZZY NEURAL NETWORK SYSTEM

Figure 3 shows the detection procedures on sub-frequencies of a CDHMM/FNN system, which consists of a CDHMM detection subsystem and a FNN detection subsystem. More specifically, the multi-subband output vector based on CDHMM is first mapped non-linearly in the vector space using FNN. Detection information is then extracted from the mapping results. After this, the components of the input vectors are extracted by making uses of non-linear mapping of the neural network (Li *et al*., 2002). Finally, components are classified based on the relevance between patterns (Say and Eng Guan, 2001).

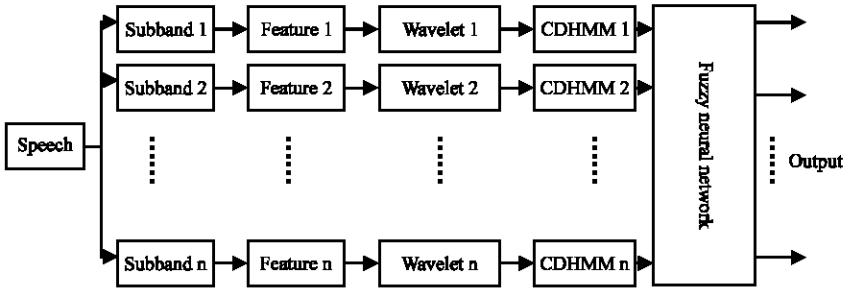

Fig. 3: Parallel subband CDHMM and Fuzzy neural network system

SIMULATION EXPERIMENT AND ANALYSIS OF RESULT

Process of system training and recognition: The system training includes wavelet transform, CDHMM and FNN training. The voice sampling rate is 8 kHz. We use 12-order MFCC and its 1st and 2nd difference parameters as the eigenvectors of the speech recognition system. During the recording, since each word may has different length, we normalize each word to 20 frames (n = 20), which means each word has 240 (n×p) eigenvectors.

The wider the sub-frequency, the more obvious are the characteristics of the class and hence the easier to do the estimation. However, capturing the local attributes often becomes problematic as the subband becomes too large. On the contrary, when the sub-frequency is narrow, it is much easier to capture the local attributes, but difficult to estimate the class, as smaller subband contains less information and the characteristic of class is not obvious. Therefore, we use the four subbands as follows: 0~778 Hz, 707~1632 Hz, 1506~2709 Hz and 2509~4000 Hz. Since quadrature transform allows elimination of signal correlation and accumulation of signal energy, signal power can be concentrated on the coefficients of certain subbands. At the same time, the coefficients of other subbands can be set to be 0 or small weights, so that noise can be suppressed effectively. Each set of data produces a CDHMM model after training which is independent to others. If the serial numbers which have the highest output probability from each model are extracted and combined with the average energy of each frame of the input speech, we get the input vector $Vi = [q_{i1}, q_{i2}, L, q_{ii}, E_i]$, in which $q_{i1}, q_{i2}, L, q_{ii}$ are the serial number of highest output probabilities of subbands x_i and the full frequency range and E_i is the average frame energy of x_i. After mapping the input vector with using FNN, we can obtain the recognition result.

We use administrative of π function fuzzy to parameters and use l as 3 according to our experience. The number of inputs on the left side of the network is 720 ($n \times p \times l$), i.e., there are 720 neural units for each neural network layer. The first and the second hidden layer have 500 and 600 units respectively. The output has 26 units, corresponding to the classification results of 26 two-symbol characters. The inputs on the right side of the network are the outputs of the left three neural networks. Since then, there are 130 input units. The hidden layer has 100 units and the output has 26 units, corresponding to the final recognition results. The hidden layer of the BP neural network uses tangent Sigmoid transfer function and the output uses linear transfer function.

The testing process: We recorded speech data for experiment in a lab environment, using CoolEdit. The word list consists of 50 computer command words. We regard the speech in such environment as pure speech. There were 20 speakers and each speaker repeated every word three times, in which 10 of them used one set of the speech as training material and the rest as recognition material. We also had 12 speeches which were recorded in different environment as noise speech.

Procedures:

- Record the speech under CoolEdit and resolve it onto 4 subbands using low-pass filter $H_1(z)$, $H_0(z)$ and high-pass filter $H_1(z)$
- Split the speech information on the 4 subbands into frames and Pre-emphasis, using VoiceBox contained in MATLAB
- Extract the eigenvectors by mfcc.m function in the VoiceBox and then remove the noise in the eigenvector using orthogonal wavelet function. The terminal is also detected to obtain speech data x_i
- Make sequential processing using CHMM and get CHMM model parameters using Viterbi algorithm. Extract the serial numbers which have the highest output probabilities from each model and combine them with the average energy of each frame of the input speech, we get the input vector $Vi = [q_{i1}, q_{i2}, L, q_{ii}, E_i]$, in which $q_{i1}, q_{i2}, L, q_{ii}$ are the

serial number of highest output probabilities of sub-frequency x_i and the full frequency range and E_i is the average frame energy of x_i
- Normalize the input vector as the input of the FNN, train the input vector using the FNN until it reaches the desired accuracy

RESULTS AND DISCUSSION

The speeches are recognized by several recognition models simultaneously. Table 1 shows the training and recognition results of pure speech and noise speech under CHMM model, multi-band CHMM/BPNN and CDHMM/FNN.

From Table 1, we can observe that:

- The CHMM has the highest recognition rate for the pure speech of non-specific person and keywords, but performs worse as the signal to noise ratio decreases
- CHMM/BPNN performs worse than CHMM for pure speech, but performs better when noise presents due to the influence of subbands. In the condition of low noise or mismatch, fusing model is considered better than the CHMM model
- In a pure speech condition, the CDHMM/FNN model performs worse than the CHMM model, but better than the CHMM/BPNN model. When noise presents, the CDHMM/FNN model outperforms the other two

The traditional CHMM model performs worse as the signal to noise ratio decreases; although the DHMM/BPNN model performs better when noise presents, has lower recognition rate for the pure speech. In fact, the two models mentioned above are not suitable for applications. The reasons of the method proposed in the paper is reasonable are as follows:

- Since, all sub-frequencies are independent to each other, they can be trained individually using different methods. This allows the system to be more adaptive and robust (Bourlard and Dupont, 1997)
- With wavelet transform, only useful information is extracted (Hu and Wu, 1999) and the calculation load is considerably reduced, thus improving on training time and detection accuracy
- Fuzzy Neural Network (FNN) is used to improve the convergence speed and to avoid local optimal (Cui, 2004)
- By making use of two recognition systems (CDHMM and FNN), the parallel subband CDHMM/FNN recognition system using wavelet processing can effectively enhance the robustness of the system in terms of noise suppression

Table 1: Comparison of three models recognition rate

CDHMM		Multi-band CDHMM/BPNN		Parallel subband CDHMM fuzzy neural network	
Pure speech	Noise speech	Pure speech	Noise speech	Pure speech	Noise speech
(%)					
97.1	64.3	89.3	72.2	89.5	79.5

CONCLUSION

An ideal speech recognition system should extract the content of the speech in various conditions correctly. In this study, firstly the Parallel subband HMM and neural network algorithm based on Wavelet analysis is discussed. Then a combined speech recognition model which constitute by CDHMM subsystem and FNN subsystem is introduced. The effectiveness of the proposed scheme is demonstrated by experiment. At the same time the recognition rate of two models (CHMM and CHMM/BPNN) and our proposed one are compared in Table 1. Finally, it is proved that the proposed system is better performance in robustness than the traditional CHMM model and CHMM/BPNN model. The recognition rate is also being improved.

In the future, we will focus on resolving the problem of the considerable training time, a single-layer neural network could instead be adopted in order to save training time.

REFERENCES

Allen, J.B., 1994. How do humans process and recognize speech? Proc. IEEE, 4: 567-577.

Bershtein, L.S. and S.M. Kovalev, 2004. Fuzzy temporal models of acoustic processes in intelligent systems of automatic speech recognition. J. Comput. Syst. Sci. Int., 6: 899-904.

Bourlard, H. and S. Dupont, 1997. Subband-based speech recognition. Proceedings of the 1997 IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 21-24, Munich, Germany, pp: 1251-1254.

Cui, M., 2004. Performance analysis and improvement countermeasure of FNN structure. Comput. Eng. Appli., 27: 99-102.

Desai, V., A. Dibazar, T.W. Berger and S. George, 2003. Using dynamic synapse based neural networks with wavelet preprocessing for speech applications. Proceedings of the International Joint Conference on Neural Networks, Jul. 20-24, IEEE Computer Society, Washington, DC., USA., pp: 666-669.

Deshpande, M.S. and R.S. Holambe, 2008. Text-independent speaker identification using hidden markov models. Proceedings of the 1st International Conference on Emerging Trends in Engineering and Technology, Jul. 16-18, IEEE Computer Society, Washington, DC., USA., pp: 641-644.

Douglas, O., 2003. Interacting With computers by voice: Automatic speech recognition and synthesis. Proc. IEEE, 9: 1272-1305.

Hu, H. and S. Wu, 1999. Application of wavelet denoising in speech recognition. J. Beijing Univ. Posts Telecommun., 3: 31-34.

Leung, K.F., F.H.F. Leung, H.K. Lam and P.K.S. Tam, 2003. Recognition of speech commands using a modified neural fuzzy network and improved GA. Proceedings of the 12th IEEE International Conference on Fuzzy Systems. St. Louis, Missouri, May 25-28, IEEE Computer Society, Washington, DC, USA., pp: 190-195.

Li, Y., B. Bai and L. Jiao, 2002. A model identification approach of non-linear systems based on fuzzy neural networks. J. Elect. Inform. Technol., 3: 355-360.

Lilia, L. and M. Sellmai, 2003. Connectionist Probability estimators in HMM Arabic speech recognition using fuzzy logic. Lect. Notes Comput. Sci., 3: 149-169.

Pulasinghe, K., K. Watanabe, K. Lzumi and K. Kiguchi, 2004.. Modular fuzzy-neuro controller driven by spoken language commands. Proc. IEEE, 1: 293-302.

Say, W.F. and L. Eng Guan, 2001. Speaker recognition using adpatively boosted classifiers. Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology, Aug. 19-22, Tencon. Singapore, pp: 442-446.

Sungwook, C., Y. Kwon and Y. Sung-ll, 2001. Denoising on adapted wavelet packets domain for robust speech recognition. Proceedings of the IEEE International Symposium on Industrial Electronics, Jun. 12-16, Pusan, South Korea, pp: 497-500.

Trentin, E. and M. Gori, 2003. Robust combination of neural networks and hidden Markov models for speech recognition. Proc. IEEE, 6: 1519-1531.

Tsung-Hui, C., L. Zhi-Quan and C. Chong-Yung, 2008. A convex optimization method for joint mean and variance parameter estimation of large-margin CDHMM. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 31-Apr. 4, Las Vegas, pp: 4053-4056.

Xinwei, L., J. Hui and L. Chaojun, 2005. Large margin HMMs for speech recognition. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 18-23, Philadelphia, pp: V/513-V/516.

Xiying, W. and D. Guozhong, 2007. A novel method to recognize complex dynamic gesture by combining HMM and FNN models. Proceedings of the IEEE Symposium on Computational Intelligence in Image and Signal Processing, Apr. 1-5, Honolulu, HI, USA., pp: 13-18.