Codebook Design Method Based on Genetic Algorithm for Chinese Continuous Digital Speech Recognition

Yanna Wei[1], Liyan Zhao
Department of Science and Engineering, North China Institute of Aerospace Engineering, Langfang, China
e-mail: weiyanna_2005@163.com

Abstract Vector quantization is one of model training and pattern matching technologies with good performance which is applied to speech recognition at present. The traditional Linde-Buzo-Gray (LBG) algorithm is easy to get the local optimal result in the process of codebook design. According to capability of getting global optimal, genetic algorithm (GA) is used to improve the quality of codebook. The new algorithm is applied to mandarin continuous digit speech recognition and implementation procedure is given in detail. The experiments show it is more effective than traditional LBG algorithm, especially in condition of lower SNR.

Keywords Speech recognition • Vector quantization • Linde-Buzo-Gray algorithm • Genetic algorithm

1 Introduction

The basic goal of speech recognition is studying out a kind of machine with hearing function which can recognize phonetic message and perform human's intention in any condition. It belongs to the field of pattern recognition and artificial intelligence. In the recent twenty or thirty years, speech recognition has been used in computer, signal processing, communication, electronic systems and automatic control field [1, 2].

Vector Quantization (VQ) is one of data compression and coding methods, and it has been used in speech coding, speech synthesis, speech recognition and speaker recognition successfully [3].

[1] Yanna Wei (✉)
Department of Science and Engineering, North China Institute of Aerospace Engineering, Langfang, China
e-mail: weiyanna_2005@163.com

In the process of codebook design, traditional Linde-Buzo-Gray (LBG) algorithm is easy to get the local optimal result and be influenced by initial codebook. GA has the capability of getting global optimal result, so a new clustering algorithm GA-L which based on GA and LBG is proposed to improve the quality of codebook in Chinese continuous digital speech recognition.

2 Codebook Optimization

In the process of VQ codebook design, the training data are classified according to certain degree of distortion criterion in LBG algorithm. This algorithm is steepest descent algorithm, operation results of objective function can only depend on the initial value of codeword, so it is difficult to guarantee that LBG algorithm is used to obtain the global minimum of the objective function [4, 5].

GA is a stochastic search algorithm, and it was put forward by American Prof. J. Holland [6, 7]. GA with reliable global search capability, it does not rely on gradient information, but simulates the process of natural evolution to search the optimal solution. It is especially suitable for processing complicated problems that are difficult to be solved by traditional search method.

Therefore, GA can be combined with VQ in order to avoid local optimal solution. The schematic diagram is shown in Fig. 1.

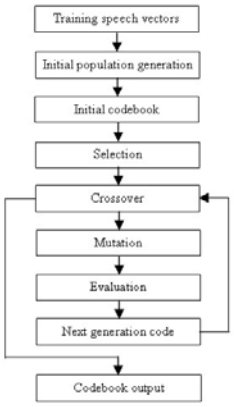

Fig. 1 The schematic diagram of VQ based on GA

The new clustering method GA-L which combines GA with LBG is used to optimize codebook in this paper.

GA don't deal with spatial data directly, it must be decoded and expressed as genotype string structure data in genetic space. The size of breeding population should be enlarged to keep the diversity of individuals, and the numerical value should be greater than 30 and less than 75 to reduce the operation time. Individual fitness function can be defined as the reciprocal of mean quantization distortion for training vector sequence. It is defined by

$$f_i = \frac{1}{\frac{1}{N}\sum_{j=1}^{N} \min_k d(X_j, Y_i^k)} \tag{1}$$

where N is the length of training vector sequence $X = \{X_1, X_2, \ldots, X_N\}$, Y_i^k is the kth codeword of individual i.

Genetic operation of speech recognition includes selection, crossover and mutation, and its effect depends on operation probability, coding method, initial population and fitness function. The operations are selection of superior individual and elimination of inferior individual from the population. It is based on the value of fitness function, the greater fitness of individual, the higher probability of being selected. Roulette wheel selection method has been adopted, in which choose probability becomes direct ratio with fitness.

The size of breeding population is defined as n and individual fitness function is defined as f_i, thus the probability of selection is given by

$$P_{si} = f_i / \sum_{j=1}^{n} f_j \tag{2}$$

Because the distance is fixed between two points in fixed-length crossover, the two points can be produced simultaneously according to crossover probability P_c. Mutation operation can prevent genetic process from local optimization, and mutation probability P_m is adopted for each gene in chromosomes. In order to recounting clustering center and renewing chromosomes, the LBG and nearest neighbor rule are implemented after crossover and mutation for all individuals. Stop condition of iteration is controlled by degree of convergence, in which evolutionary algebra is used to control heredity.

3 Chinese continuous digital speech recognition based on GA-L

Chinese continuous digital speech recognition based on GA-L includes preprocessing, feature extraction, model training and distortion measure, which is shown in Fig. 2.

3.1 Preprocessing

Preprocessing includes anti-aliasing filtering, sampling, quantization, denoising, selection of speech recognition unit, endpoint detection, pre-emphasis, multiplying window and framing.

After processed by window function, the length of which is N, speech signals are detected according to frames, the time of each frame is about 10ms~30ms, the short-term energy of the nth speech signal $x_n^2(m)$ is given by

$$E = \sum_{m=0}^{N-1} x_n^2(m) \tag{3}$$

E is a measurement function of signal amplitude. To avoiding error caused by square in (3) as too large or too small sampling value, which is expressed as amplitude function M defined as

$$M = \sum_{m=0}^{N-1} |x_n(m)| \tag{4}$$

Short-term average ZCR is the times of passing zero level per frame of signals. It is given by

$$ZCR = \frac{1}{2} \sum_{m=0}^{N-1} |sign[x_n(m)] - sign[x_n(m-1)]| \tag{5}$$

where sign[$x(n)$]=1($x(n)\geq 0$) or sign[$x(n)$]=-1($x(n)<0$).

The sampling rate of speech data is 11025Hz, frame size is 256 points and frame shift is 80 points, Signal Noise Ratio (SNR) is less than 6dB. Simulation of speech signal 'si' by MATLAB in laboratory environment and extraction of waveform, average amplitude and zero-crossing of this speech, as shown in Fig. 3. Obviously, the characteristics of unvoiced consonants are less energy and higher zero-crossing rate.

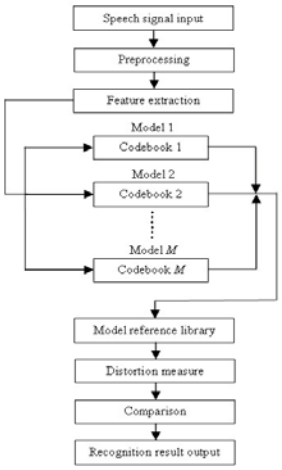

Fig. 2 The schematic diagram of chinese continuous digital speech recognition based on GA-L

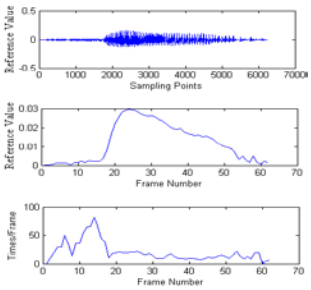

Fig. 3 Waveform, average amplitude and zero-crossing of speech signal 'si'

3.2 Characteristic Parameter Extraction of Speech Signal

Mel Frequency Ceptral Coefficient (MFCC) accords with the characteristics of human auditory system, and it is more stable in channel with noise [8]. The relationship between Mel frequency and the actual frequency can be shown as

$$Mel(f) = 2595\lg(1+ f/700) \tag{6}$$

Logarithm is operated for $m(l)$, then Discrete Cosine Transform (DCT) is performed, last MFCC is obtained. The formula of calculating MFCC is given by

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

12 dimensions MFCC and 12 dimensional first-order differential MFCC are adapted in speech recognition.

3.3 Model Training

Good codebook must be designed before recognition, which needs a lot of input signal. The process of codebook design is a “training”or “learning”, in which clustering algorithm is applied [9]. According to certain distortion criteria, training data are clustered and pattern library is formed. Recognition based on VQ is a kind of non-parameter model, each codeword in the pattern library corresponds to one codebook vector.

GA-L based on GA and LBG is used to improve the quality of codebook. Selection, crossover and mutation of individual are carried out in GA, which overcomes the defects of local optimization in LBG to get the global optimal solution.

The flow diagram of GA-L is shown in Fig. 4. Genetic manipulation is used and the fitness of individuals was calculated, on the basis of which the LBG algorithm is implemented to recounting clustering center.

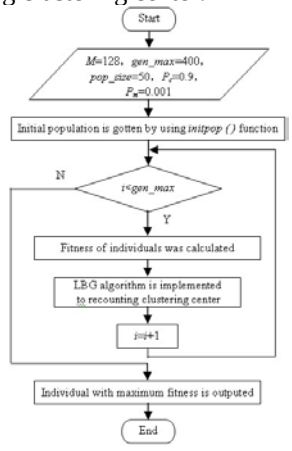

Fig. 4 The flow diagram of GA-L

The codebook distribution of speech signal ‘si’ is obtained by using GA-L algorithm, as shown in Fig. 5. The waveform and image of codebook matrix are are shown in Fig. 6. In addition, the waveform and image of recognition results matrix are shown in Fig. 7.

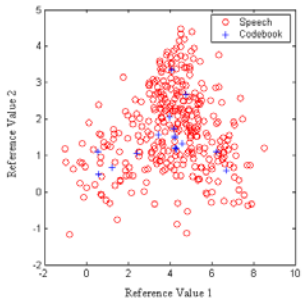

Fig. 5 The codebook distribution of speech signal ‘si’

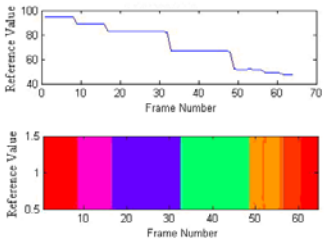

Fig. 6 Waveform and image of codebook matrix

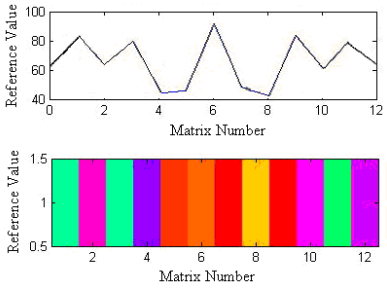

Fig. 7 Waveform and image of recognition results matrix

3.4 Distortion Measure

According to certain rules (such as some kind of distortion measure), the similarity between input feature vector and inventory model is calculated (such as matching distance) [10]. It is given by

$$d(X_n, Y_l^i) = \sum_{k=1}^{K} (X_{nk} - Y_{lk}^i)^2 \tag{8}$$

Where $d(X_n, Y_l^i)$ is the distance between test vector X_n and code vector Y_l^i, Y_l^i is the kth component of the lth code vector in the ith codebook.

3.5 Results and Comparisons

In order to confirm the universality of improved method, 10 groups of detection speech are adopted in this experiment. It includes spirant, rhonchus, lateral, semi-vowel, and so on. Model reference library is recorded by 20 people (10 male, 10 female), 3000 speech data are used for model training.

GA-L algorithm and LBG algorithm are used to design codebook separately. The experimental results are shown in Table 1, in which the "recognition rate 1" shows the results of using GA-L and "recognition rate 2" shows the results of using LBG.

Table 1 The experimental results

Length of speech data	Recognition rate 1	Recognition rate 2
1	99.7%	96.6%
2	97 .5%	93.2%
3	94.7%	90.0%
4	91.4%	87.0%
5	87.4%	83.8%
6	84.3%	80.9%
7	80.5%	78.3%
8	77.5%	75.1%
9	74.2%	72.7%
10	72.1%	70.4%

The experiment results show that "recognition rate 1" is higher than "recognition rate 2", the average recognition rate can be improved by 3.16% or so by using the improved algorithm, which demonstrates the effectiveness of GA-L.

4 Conclusions

The improved algorithm GA-L avoids falling into local optimum in the process of codebook design for Chinese continuous digital speech recognition. GA-L makes full use of global search of GA and fast convergence of LBG. It is better than the traditional LBG. Experimental results show the efficiency of the new algorithm.

Acknowledgements This work is supported by Langfang science and technology support project (No.2012011012 and No.2012011024).

References

1. AhmedB and HolmesW H, “A voice activity detector using the square test,” Proc ICASSP’2004, 2004, pp.625-628.
2. Chao Huang, Tao Chen, and Eric Chang, “Speaker selection training for large vocabulary continuous speech recognition,” in Proceedings International Conference on Acoustics, Speech, and Signal Processing, Orlando, 2002, pp. 609-612.
3. Gao Chang, LI Haifeng and Ma Lin, “Content-based compressive sensing for speech signal,” Journal of Signal Procesing, vol. 28, pp. 851-858, June 2012.
4. K.Lamia and M.Arnaud, “Towards improving speech detection robustness for speech recognition in adverse conditions,” Speech Communication, vol. 40, pp. 261-266, March 2003.
5. N.T.Lay, F.W.Say, and D.Silva, “Robust endpoint detection and energy normalization for real-time speech and speaker recognition,” IEEE Trans. On Speech and Audio Processing, vol. 10, pp. 146-157, March 2000.
6. Sawit Kasuriya, Chai Wutiwiwatchai, and Varin Achariyakulporn, “Comparative study of continuous hidden markov models and artificial neural network on speaker identification system,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 9, pp. 673-683, June 2001.
7. Tian Ye, Wu Ji and Wang Zuoying, “Fuzzy clustering and bayesian information criterion based destimation for robust voice activitydetection,” IEEE Trans Acoustics, Speech, and signal processing, vol. 2, pp. 444-447, January 2003.
8. Wen Gao, Yiyong Ma, and Jiangqin Wu, “Sign language recognition based on HMM/ANN/DP,” International Journal of Pattern Recognition and artificial Intelligence, vol. 14, pp. 587-602, May 2000.
9. Ye Lei, Sun Linhui and Yang Zhen, “Endpoint detection algorithm based on cepstral distance of compressed sensing measurements of speech signal,” Journal of Signal Procesing, vol. 27, pp. 67-72, January 2011.
10. Zhu Xuan, Chen Yining and Liu Jia, “A novel efficient decoding algorithm for CDHMM based speech recognization,” IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, 293-296.