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 12 (20): 5661-5664, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.5661.5664

Research of Verification-Mode Voice Recognition Based on UBM-GMM

[1,2]Fang Zhi-Gang, Sun Chao, [2]Xu Jie
[1]Institute of Electronic Circuit and Information System, Zhejiang University, 310007,
Hangzhou, Zhejiang Province, China
[2]Zhejiang University City College, 310015, Hangzhou, Zhejiang Province, China

Abstract: This study proposes a Verification-Mode voice recognition system which uses in IC card. Improvements have been made both in the two periods of voice recognition: on the one hand, a new method of endpoint detection is used in voice pretreatment and doing cepstrum filtering for coefficients in the MFCC extraction; on the other hand, use UBM-GMM algorithm in the process of template matching and make it better to the verification mode. Experiment with putting the user's voice data into IC card to do the user verification and the result shows that this mode eliminates the dependence of voice text and improve the speed and precision of the identification.

Key words: Verification-mode, endpoint detection, UBM-GMM algorithm

INTRODUCTION

According to the degree of dependence on text, Verification-Mode voice recognition system can be divided into two kinds: Text-dependent and Text-independent. The Text-dependent system requires users to pronounce the specified content, everyone's voice model will be established accurately one by one and users must also pronounce the specified content when doing identification, hence, it can perform good recognition effect, but the system needs users' cooperation. On the contrary, the Text-Independent system don't set the voice content, so users to use more convenient and the only lack is modeling difficulty and large storage requirement.

This study presents a voice recognition system based on GMM, do the Text-Independent voice recognition experiment on the database in size of 25 people (40 audio/people), feature selection uses the Mel Frequency Cepstrum Coefficient (MFCC) and using the training algorithm based on Universal Background Model-Gaussian Mixture Model (UBM-GMM) (Reynolds and Rose, 1995), which improves the speed and precision of the identification to a certain extent.

GAUSSIAN MIXTURE MODEL

GMM describes the voice characteristics distribution with several multidimensional gaussian probability density linear combination and when do voice recognition, the eigenvector sequence of test voice go through the model in turn. Then, calculate the probabilistic likelihood score of the matching output and take the total output probability of all frames as the judgment basis.

For a eigenvector sequence X = {x_1, x_2,..., x_T}, where x_t (T = 1,.... ,T) mean d dimensional eigenvector of discrete time t ∈[1., 2, ..., T]. The total output probability is:

$$P(X|\lambda)=\left[\prod_{t=1}^{T}p(x_t|\lambda)\right]^{1/T} \quad (1)$$

Where:

$$p(x_t\,|\,\lambda)=\sum_{i=1}^{M}C_i g(x_t|\mu_i,S_i) \quad (2)$$

where, M is the number of gaussian mixture model, which called mixedness; C_i, i = 1, 2, 3, ..., M is the weight value of single gaussian distribution mixture and it satisfies:

$$\sum\nolimits_{i=1}^{M}C_i=1$$

The i-th single gaussian distribution is:

$$g(x_t|\mu_i,S_i)=\frac{1}{(2\pi)^{d/2}\,|\,S_i\,|^{1/2}}\exp\{-1/2(x_t-\mu_i)^T S_i^{-1}(x_t-\mu_i))\} \quad (3)$$

This is a normal distribution probability density function, μ_i is the d dimensional mean vector of i-th single

Corresponding Author: Sun Chao, Institute of Electronic Circuit and Information System, Zhejiang University, 310007, Hangzhou, Zhejiang Province, China gaussian distribution, Σ_i is its d×d covariance matrix. GMM's training is based on Maximum Likelihood (ML), to a set of training vector, the parameters of ML model are getting with iteration EM (Expectation-Maximization) algorithm (Zhang *et al*., 2003). Each iteration has two steps, E-step estimates the no-study data's distribution according to the known study sample and the parameters of the before; M-step estimates the ML model parameters and repeat until the local maximum, under the assumption that E-step distribution is correct.

REALIZATION OF VOICE CONFIRMATION PROCESS

Voice confirmation process is mainly made up of two parts: Extraction of voice feature and selection of model classifier. Feature extraction uses the method of MFCC, since it has strong antinoise ability and the auditory characteristics accord with human ear. And the sample training chooses GMM to make model, because it has good recognition effect and does't depend on voice text. Considering that the actual samples of voice signal contain large amount of redundant information (quiet period and low amplitude background noise), it will be low recognition speed and accuracy if the system process the samples without detection and will ultimately affect the recognition performance of the system. Hence, the author introduces a kind of high efficient voice-signal endpoint detection mechanism (Canny, 1986) on the basis of traditional one to improve the system performance.

Figure 1 shows the changes of voice signal before and after treated with endpoint detection. It can be seen that the 45000 unit length of original voice change to 10000 unit length of existing voice signal, which saves the processing time for computer and improves the speed of voice feature parameters processing. Meanwhile, it eliminates the noise segments and improves the recognition accuracy and the threshold is adjustable, so it has flexible adaptability.

Firstly, the voice signal will do normal de-noising processing and endpoint detection and then, it adds window frames. After that, the MFCC coefficient extraction of every frame (Sambur, 1975) has 5 steps: Fast Fourier Transform (FFT), Mel frequency filtering, natural logarithm operation, Inverse Discrete Cosine Transform (IDCT) and cepstrum filtering. What is really needed to demonstrate that the purpose of the Mel cepstrum filtering is making the spectrum envelope and harmonic information of signal more clearly. The length L of cepstrum filter should take two-thirds of IDCT length, filtering form as follows:

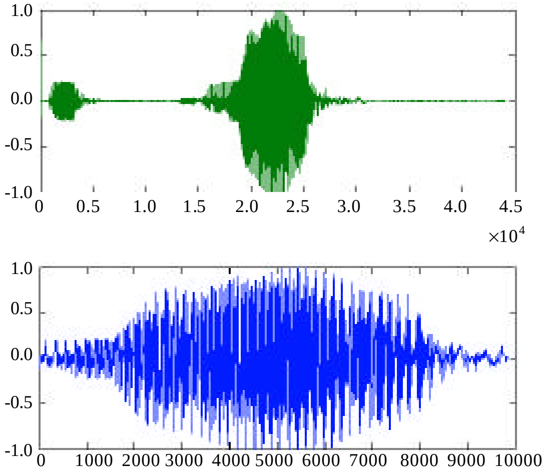

Fig. 1(a-b): Voice before detection, (b) Voice after endpoint detection

$$c_{mel} = c_{IDCT}(1 + \frac{L-1}{2}\sin(\frac{\pi n}{L-1})) \tag{4}$$

where, c_{mel} is Mel cepstrum coefficient, c_{IDCT} is the coefficient after IDCT.

WORKFLOW OF SYSTEM

Assume that the probability of eigenvector sequence X generated by the voice S is:

$$\Pr(\lambda_s|X) = \frac{p(X|\lambda_s)\Pr(\lambda_s)}{p(X)} \tag{5}$$

where, p $(x|\lambda_s)$ is the probability density of sequence X produced by model λ_s; Pr (λ_s) is the prior probability, generally assume that Pr (λ_s) are equal; p (x) is the prior probability of X from any voice.

The purpose of voice recognition is to determine whether the given period testing voice is the original voice, in the case of a known target voice model. It can be seen as a hypothesis testing problem, that is to say, determine the test voice is made by real voice (H_0) or pretender (H_1), hence, Higgins *et al.* (1991) proposed the method of use Likelihood Ratio to express confirm score in voice recognition. Likelihood Ratio is defined as the ratio of the probability that testing voice made by real voice model and the probability that testing voice made by pretender voice model. Using Pr $(\lambda_{tar}|X)$ to express real voice model, Pr $(\lambda_{imp}|X)$ to express pretender voice model, so the Likelihood Ratio is:

$$L(X) = \frac{\Pr(\lambda_{tar}|X)}{\Pr(\lambda_{imp}|X)} = \frac{p(X|\lambda_{tar})}{p(X|\lambda_{imp})} = \frac{\prod_{t=1}^{T} p(x_t|\lambda_{tar})}{\prod_{t=1}^{T} p(x_t|\lambda_{imp})} \tag{6}$$

There assume eigenvectors x_1, x_2,.... X_T are independent of each other. Usually evaluate the logarithm of Eq. 6 and then get the logarithmic likelihood score by doing average. Meanwhile, in order to reduce the pronunciation time's effect on Likelihood Ratio, we take the form of time normalized logarithmic Likelihood Ratio:

$$\begin{aligned}\Lambda(X) &= \log(L(X)) = \log\left(p(X|\lambda_{tar})\right) - \log\left(p(X|\lambda_{imp})\right) \\ &= \frac{1}{T}\left(\sum_{t=1}^{T} p(x_t|\lambda_{tar}) - \sum_{t=1}^{T} p(x_t|\lambda_{imp})\right)\end{aligned} \tag{7}$$

and then compare the LR with threshold to determine whether accept H_0 or refuse H_0. Because it is voice recognition, we can't make model for all the pretenders, but:

$$\sum_{t=1}^{T} p(x_t|\lambda_{imp})$$

can be experience set. Hence, voice recognition's judgment criterion is as follows:

$$D(X) = \begin{cases} \text{ACCEPT}, & \Lambda(X) > \text{Threshold} \\ \text{REFUSE}, & \Lambda(X) \le \text{Threshold} \end{cases} \tag{8}$$

RESULTS OF EXPERIMENT

Experiment designed to study the effect of GMM hybrid number and test voice time to recognition performance, which in the voice recognition system based on GMM. We got the data from ORATOR Expressive voice database, for this database has 117 different versions to the same 8 paragraph by 13 professional actors and 14 common people. Experiment collected 25 samples deducing by different people and sampling at the frequency of 44.1 Khz, then divided each sample into 40 sections, each section of 0.2 and 0.52 sec as a contrast experiment; the whole sample before sectioning as a training, the total duration is about 25 sec.

Figure 2 describes the ROC curve when test voice is just 0.2s, GMM hybrid number M respectively take 32, 64, 128, 256, 512. Intersections of ROC and dotted lines are EER. As can be seen from Fig. 2, when M is 256, EER is lowest, just 4.52%; but when M is 512, EER increases to 6.52%. The experiment results showed that, 25s voice signal not enough to train bigger GMM and small GMM can't make full use of the information provided by the training data.

Figure 3 is depicting the ROC curve when test voice is 0.52s, GMM hybrid number M respectively take 32, 64, 128, 256, 512. Intersections of ROC and dotted lines are EER. As can be seen from Fig. 3, when M is 256, EER is lowest, just 3.88% and when M is 512, EER increases to 4.99%. Test results are similar to the above, that is, 25s voice signal not enough to train bigger GMM and small GMM can't make full use of the information provided by the training data. However, test voice duration increase from 0.2s to 0.52s and the performance get improvement. For example, when M = 256, EER decrease from 4.52% to 3.88%, decreased by 14.2%, but at the cost of computing time and necessary information storage capacity is increased.

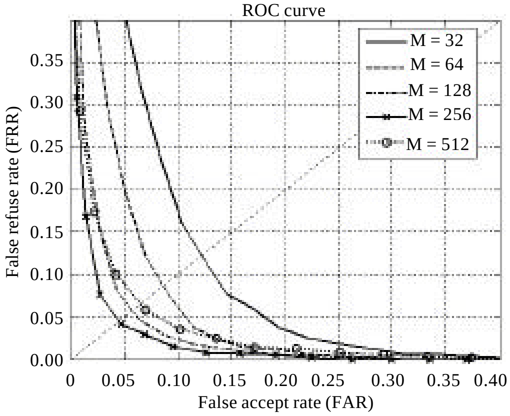

Fig. 2: ROC curve (0.2s)

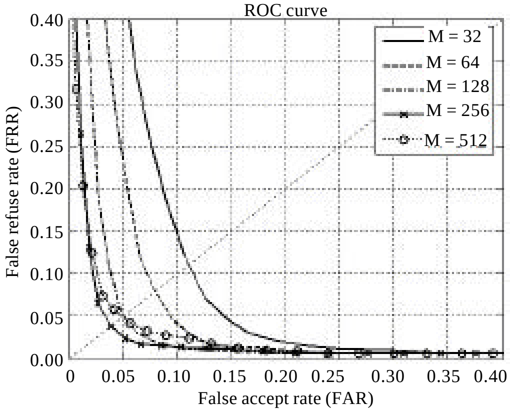

Fig. 3: ROC curve (0.52s)

CONCLUSION

This study introduces voice recognition algorithm based on GMM which is Text-Independent and applied it into verification mode. Experiment results showed that when the test voice was only 0.52s and GMM hybrid number M = 256, its EER was 3.88%. After endpoint detection, only to extract the useful information, thus reduced the complexity of voice processing. However, in terms of the system security, such as individual characteristic information encryption, remains to be further discussed and the relevant studys will be published in succession.

ACKNOWLEDGMENTS

Our Research Project was fully sponsored by Bureau of Hangzhou city science and technology project, Hall of Zhejiang province science and technology public welfare project and Zhejiang University City College cross research foundation project with grant number 20120633B32, 2012C21038 and J-13027. Thanks for the sponsorships.

REFERENCES

Canny, J., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell., 8: 679-698.

Reynolds, D.A. and R.Ñ. Rose, 1995. Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. Speech Audio Process., 3: 72-83.

Higgins, A.L., L. Bahler and J. Porter, 1991. Speaker verification using randomized phrase prompting. Dig. Sig. Process., 1: 89-106.

Sambur, M., 1975. Selection of acoustic features for speaker identification. IEEE Trans. Acoustics Speech Signal Process., 23: 176-182.

Zhang, Z., C. Chen, J. Sun and K.L. Chan, 2003. EM Algorithms for Gaussian mixtures with split-and-merge operation. Patt. Recog., 36: 1973-1983