Multi-stages decision voice print recognition in telecommunication

Fa Jiang Ma [1], Lan Tian[1] & Xiao Shan Lu [1,a]

[1]*school of information Sci. and Eng., Shandong Univ., jinan, China*

Abstract. Using the cepstrum and pitch parameter, a VQ voice print recognition (i.e. speaker identification) method was realized and used in telecommunication. In this paper, we analyzed the characteristic of long-time LPCCEP and the pitch parameter for text-independent speaker identification. Considering the different contribution of each LPC cepstrum component, the weight was added to modify the distortion measurement. Based on the improved vector quantification (VQ), we constructed the speaker individual model and the system was training shorter, the response real time and the data memory smaller. A simple pitch measurement was introduced and applied in multi-stage decision to raise system robustness. The test results show that using about 10 seconds voice, in 60 speakers set, the correct identification rate is above 95%.

Keywords: Voice print recognition; Vector quantification(VQ); LPC cepstrum; Pitch.

Introduction

Recently, voice print recognition technology has caused widespread interests for its convenience, economic, and veracity. Particularly in identity automatic identification based on telecommunication, as one of the human biology characteristics, voice print is advantaged and irreplaceable.

Voice print recognition is realized when speakers provide the system with speech materials, considering whether speech materials is fixed or not, it is divided into two cases: text-dependent and text-independent. The later is more practical because of less limitation, and can prevent from voice bootleg in use. This paper introduces a rapid and accurate recognition method, which analyzes and processes readily, then tells the speaker form all set in high correct rate, which based on an arbitrary speech, when speech materials are unstable and irregular. We will discuss this method in the following aspects: feature parameter selection, recognition modeling, system structure, and test experiments. This technology has been applied to some business system,such as: SpeakerKey of ITT company, VoiceGuardian of Keyware company and so on[1]. Several problems should be solved in voice print recognition technology, such as, the unsteady recognition capacity because of voice drift[2], and the extracted information lacking of emotion factor[3].

Speech Feature Analysis

Feature parameters extraction is crucial in voice print recognition. Pitch is an obvious acoustic feature. The pitch periods of different speakers vary not so much, could be imitated easily, and change a lot over time. Therefore it has a poor effect to recognize, when using pitch parameters only.

Comparatively speaking, vocal tract features change a little over time, and can't be imitated easily. And it is the more important physiology difference of speakers' personality, and the spectrum envelope of speech is the acoustic performance of vocal tract. In comparison to pure unvoiced sound, voiced sound has a high frame loudness and high SNR, and contains more speakers' information. Applying linear prediction analysis technology, vocal tract and voice source feature can be separated effectively. Particularly, we analyze some typical vocal tract parameters, including LPC, PARCOR and LPC cepstrum coefficient. LPC cepstrum coefficient performs best in speaker recognition, and is easy to calculate. Through long-time average processing to LPC cepstrum coefficient , speech content information would be weakened, but vocal tract characteristic is almost invariant. Independent of vocal tract, pitch information can be applied to raise the system robustness.

Feature Parameter Extraction[4]

According to the above mentioned analysis, applying linear prediction analysis method, we extract long-time LPC cepstrum vector and pitch period, and the extraction algorithm is shown in the figure 1. In the minimum phase condition of all-pole vocal tract model, the solution formula of short-time LPC cepstrum can be deduced:

$$\hat{h}(1) = a_1 \quad \hat{h}(n) = a_n + \sum_{k=1}^{n}(1-\frac{k}{n})a_k\hat{h}(n-k) \quad 1 < n \le p$$

$$\hat{h}(n) = \sum_{k=1}^{n}(1-\frac{k}{n})a_k\hat{h}(n-k) \quad n > p \qquad c(n) = 1/2[\hat{h}(n) + \hat{h}(-n)] \tag{1}$$

[a]Corresponding author：2281176576@qq.com

Where a_k (k=1,2,...,p) is a predictive coefficient, and can be solved by Durbin recurrence method from speech signal $S(n)$, and p is the predictive order. $\hat{h}(n)$ is the LPC complex cepstrum, and $c(n)$ is the cepstrum. The lower part of the cepstrum decays rapidly as n increases, therefore a p-order cepstrum vector far less than frame width(N) could express the vocal tract characteristics of a voice frame, and P values 8 to 16 typically. Applying the adaptive endpoints detection method, to detect voiced sound, we average every 10 frames to consist a frame of long-time LPC cepstrum vector.

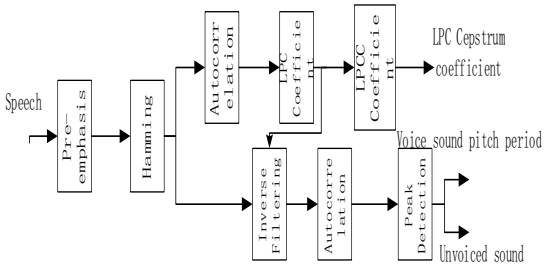

Fig. 1 The block graph of features extraction

In the quantitative analysis, when P values 16, the each component of 16-orders LPC cepstrum coefficient contributes differently, and 6-order to 16-order components contribute much more significantly and obviously. Using this phenomenon to modify the distortion measurement formula, the system response speed of speaker recognition would be raised.

Build Recognition Model on VQ[5]

In recent 20 years, Vector Quantification is a well-developed effective compression coding technology. Considering its great classification capacity, the statistical information of long time speech characteristic parameters would be quantified integrally to distinguish different speakers. Simultaneously, data would be compressed effectively , speech segmentation and time warping problem that hard to deal with would be avoided. VQ speaker identification system is shown in Fig.2.

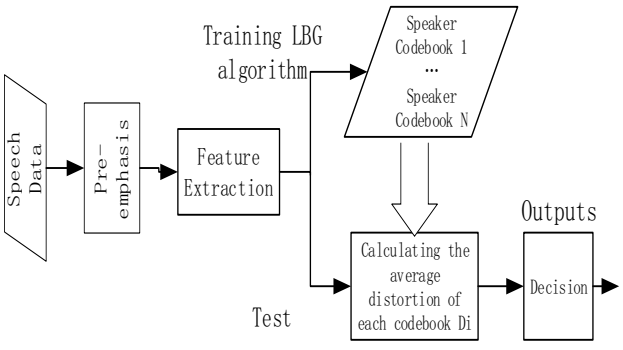

Fig.2 VQ speaker identification system

We regard the speaker personality characteristic as a information source, and model it using VQ technology. We use long time LPCCEP parameters as the information source characteristic of speakers and built N different code books corresponding to N speakers. Owing to the massive data and the uniform distribution of training vectors in feature space, the original code book is selected randomly to save time. Iteration steps are as follows:

1. Initialization: The number of given code word of the code book is L, distortion threshold ε>0, randomly select original code book $Y_L^{(0)}$, X_k(k=0,1,...,K) is the training sequence, at first iterations n=0, average distortion D_{-1}=∞;
2. According to the given code books $Y_L\{Y_i, i=1,2,...,L\}$, we use training sequences $X_k \in S_i$ make $d(X_k, Y_i)<d(X_k, Y)$ ($Y \epsilon Y_L$), to get the optimum boundary $S_i^{(n)}$, then calculate the average distortion of the training sequence in this optimum zone boundary:

$$D^{(n)} = \frac{1}{K}\sum_{k=0}^{K-1} \min_{Y \in Y_M}^{d(X_k, Y)} \quad (2)$$

3. If $\frac{D^{(n-1)} - D^{(n)}}{D^{(n)}} \le \varepsilon$, Y_L is the final code book and $S_l^{(n)}$ is the designed zone boundary, and the iteration is end. Otherwise, continue.
4. Calculating these L zones' centers of form, and using these L centers to make up the new code book $Y_L^{(n+1)}$ of the (n+1) time iteration, (n+1) replace n, and turn to step 2.

Considering the different contributions of each component of LPCCEP parameters, we should develop the average distortion measurement formula, that is, to weigh euclidean distance,

$$d(X,Y) = \sum_{i=1}^{p} \rho(i)(x_i - y_i)^2 \quad (3)$$

Where p=feature dimension, $\rho^2(i)=1/\sigma^2(i)$, and $\sigma(i)$ is the variance mean of each dimension component. The improved distance formula reflects the recognition capacity of each component adequately, thus raises the system response speed. Test speech feature vector sequence X_k(k=1,2,...,K) and all speakers' code books should calculate their average quantification distortion,

$$D_i = \frac{1}{K}\sum_{k=1}^{K} \min_{1 \le j \le L}^{d(X_k, Y_j^i)} \quad (4)$$

Then choose the smallest one or some among D_i as the result and make further judgment.

Multi-features and Multi-stages Decision Method

We adopt LPC cepstrum based on VQ to make the first decision in the identification process, then use Multi-stages decision strategy combining pitch characteristic to make further decision. In the first

decision, LPC cepstrum vectors of the test speech pick up M speakers which have the minimum distortions, M<20%*N generally. Then we propose the pitch distortion measurement formula,

$$D_i^P = \alpha_1 |\mathrm{P} - \mathrm{P}_i| + \alpha_2 |\sigma - \sigma_i| \quad 0 < \alpha_1, \alpha_2 < 1$$

(5)

Where P,σ are the pitch mean and variance of the test sample, α_1,α_2 value 1 usually. For M speakers' templates, combining this two features, the combined measurement formula is defined as the final identification result.

$$D_i^* = \beta * D_i + (1-\beta) * D_i^P \quad 0 < \beta < 1$$

(6)

If the distortion with the optimum code book is less than the specified threshold, then the corresponding speaker is the one to identify, otherwise the test speech is the sample outside the set.

Identification Experiments and Results

The original speech data was collected by telephone recording card or multimedia sound card, and the speakers consisted of 60 persons speaking different dialects, including 36 males and 24 females. All the samples are 5 to 30 seconds long, sampled in 4kHz with 16 bit resolution and the telephone speech bandwidth was within 4 kHz generally. The digital filter transmission function is $H(z)$=1-0.96z^{-1} for pre-emphasis, then the speech was processed by Hamming window, and the window width is 30 ms. The pre-processing adopted adaptive multi threshold zero-crossing rate and short-time energy to detect steady voiced sound, to calculate LPC coefficient one frame by one frame, then to acquire pitch period and each order LPCCEP coefficient (p=16) as feature vectors. Using above multi-stages decision to test, the results were shown in Fig.3. The experiment results indicated that, test length, code book size L and false recognition rate had an obvious dependence relationship, that is, the system recognition rate is higher when test speech is longer or L is larger. When L values 64 and the speech length is 10 seconds, the recognition rate is above 95%.

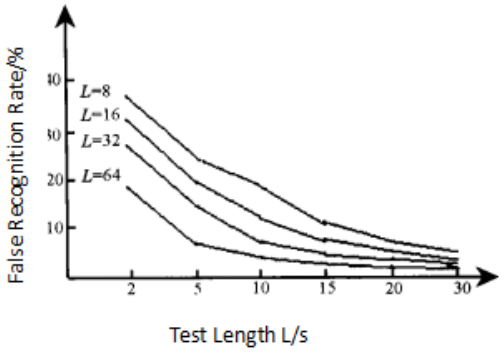

Fig.3 the results of test

Conclusion and Suggestion

In the speaker identification system based on VQ, adopting the pitch and cepstrum combined method, using weighting distortion measurement considering unbalance contribution of different components, therefore the founded voice print identification system is more complete than the cepstrum only system. In addition, this paper puts forward the decision strategy of voice tract characteristic identification firstly and then pitch feature recognition lastly, which is effective to avoid imitative testers to some extent, and it is also an important measure to raise the system robustness.

This method will be improved in the following aspect, when using LBG algorithm to cluster, the splitting method is time-consuming, and the random method is not precise enough. The system would be more practical if a precise and time-saving cluster algorithm could be used.

Acknowledgment

This paper, the research was sponsored by the National Nature Science Foundation of China (Nos.11474185, 61271453) and the Fundamental Cross-decipline Research Foundation of Shandong University(No. 2015JC029).

References

[1] J.P.Campbell. JR. Speaker Recognition A Tutorial, Proceedings of the IEEE, VOL.85, N0.9, September 1997.

[2] J.D.Woodward, Biometrics: Privacy's Foe or Privacy's Friend? Proceedings of IEEE, VOL.85, No.9, September 1997.

[3] ZHAO Li, JIANG Chun-hui, A Study on Emotional Feature Analysis and Recognition in Speech, ACTA ELECTRONICA SINICA, VOL.32, No.4, April 2004.

[4] Xingjun Yang, Huisheng Chi. Speech signal processing[M], Electronic Industry Press, Beijing, 1995.

[5] GISH H, SCHMIDTM. Text-independent speaker identification [J]. IEEE Signal Processing Magazine,1994,11(4):18-32.