A HIGH-PERFORMANCE TEXT-INDEPENDENT SPEAKER IDENTIFICATION SYSTEM BASED ON BCDM

Jin Qin, Si Luo and Hu Qixiu

Computer Science Department

Tsinghua University, Beijing 100084, P.R.China

Tel. +86 10 62786910, +86 10 62784141, FAX: +86 10 62771138,

E-mail: {jin,siluo}@sp.cs.tsinghua.edu.cn, xxs-dau@mail.tsinghua.edu.cn

ABSTRACT

This paper describes a Text-Independent Speaker Identification System of high performance. This System includes two subsystems, one is the close-set speaker identification system; the other is the open-set speaker identification system. In the implementation of the Text –Independent Speaker Identification System we introduce an advanced VQ method and a new distance estimation algorithm called BCDM (Based on Codes Distribution Method). In the close-set identification, the Correct Recognition Rate is 98.5% as there are 50 speakers in the training set. In the open-set identification, the Equal Error Rate is 5% as there are 40 speakers in the training set.

1. INTRODUCTION

Speaker recognition is an important branch of speech processing. It is the process of automatically recognizing who is speaking by using speaker-specific information included in speech waves [1]. It is getting more and more attentions due to its practical value. Speaker recognition can be classified into speaker identification and speaker verification. Speaker identification methods can be classified into text-dependent and text-independent methods. This paper only concerns text-independent speaker recognition; On the other hand, there are two cases in speaker identification which are called “close-set” identification and “open-set” identification. In close-set speaker identification system it will choose a speaker in the training set who most matches the unknown speaker as the identification decision without regarding whether he/she is in the training set or not. While in open-set speaker identification system the reference model of the unknown speaker may not exist in the training set, thus an additional decision alternative (the unknown does not match any of the models in the training set) is required. Open-set speaker identification can be applied to a lot of cases such as criminal investigations, so it is more practical in reality than the close-set speaker identification. And it is more difficult than the close-set speaker identification problem for it is required to give an additional decision alternative (Rejection Decision).

We apply an advanced VQ algorithm and a new distance estimation algorithm called BCDM (Based on Codes Distribution Method) in our system. The equal error rate of our system has decreased considerably due to our new methods as above.

2. CLOSE-SET SPEAKER IDENTIFICATION SYSTEM

Normally speaker characteristics are involved in his/her long-term utterances. A lot of training data is required in order to get enough speaker characteristics. If all the characteristics of training data are kept, A lot of space and time cost is needed. It is impractical in real-time applications.

VQ algorithm is used in many text-independent speaker identification applications. [2] Its main point is to compress the speech data by using Vector Quantifying (VQ) technique. That is to create a code- model for every speaker, and to use a kind of distance estimation method to estimate the similarity between the unknown speaker and the trainers, then to give the decision according to the minimum distance.

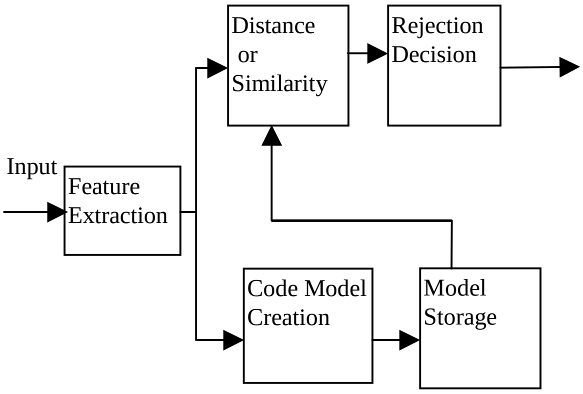


Figure 1: Block diagram of a VQ-Based close-set speaker identification system

In fact in speaker identification not all frames of the speaker’s speech data are useful to express the speaker’s characteristics. So we think maybe in recognition stage, we can only compute the distance between the code-model and the speech data that are useful for expressing the speaker characteristics. We think in feature distribution, every code-word’s central vector is the most important vector to express the speaker’s characteristics.

So the vector which has almost equal distance to no less than two code-words should be ignored.

In classical VQ algorithm the following formula is used to estimate the distance between the unknown speech and the VQ code-model:

$$D_j = \frac{1}{T}\sum_{k=1}^{T} \min_{1 \le p \le M} d\left(a_k, b_p^j\right) \qquad (1)$$

D_j is the distance between the unknown speech data and every trainer's code-model. M is the total number of code-words. a_k (1• k• T) is the feature of one frame of the unknown speech data. T is the total number of frames.

b_p^j is the pth vector of the jth speaker. And the

$$d(x,y) \qquad (2)$$

is the distance estimation formula.

There are two kinds of distance estimation formula in classical VQ algorithm.

Absolute distance formula:

$$d(x,y) = \sum_{i=1}^{p} |x_i - y_i| \qquad (3)$$

Euclid distance formula:

$$d(x,y) = \sum_{i=1}^{p} (x_i - y_i)^2 \qquad (4)$$

p is the feature dimension.

In training VQ code-model, both the central vector of every code-word and the average distance between every code-word's internal vectors and the code-word's central vector are recorded. In recognition, not all frames of the unknown data are used to give the decision. We apply a data selection method. That is: if the distance between a frame of testing speech data and its nearest code-word's central vector exceeds several times (we call it ITH -- ignoring threshold) of the code-word's average distance, the frame will be ignored.

If the data selection method is applied, We can change the function into:

$$D_j = \frac{1}{C}\sum_{k=1}^{T} \min_{1 \le p \le M} d\left(a_k, b_p^j\right) \qquad (5)$$

C is the number of the frames that are calculated, and C plus the number of the speech data frames that are ignored equals to T.

$$d(x,y) = \begin{cases} \sum_{i=1}^{p} |x_i - y_i| & \sum_{i=1}^{p} |x_i - y_i| \le ITH * A\left(C_p^j\right) \\ 0 & \sum_{i=1}^{p} |x_i - y_i| > ITH * A\left(C_p^j\right) \end{cases} \qquad (6)$$

C_p^j is the pth code-word of the jth speaker. $A\left(C_p^j\right)$ is tis average distance.

We did the following experiment based the hypothesis as above. The training set includes 33 speakers, 20 of them are males and 13 of them are females. For each speaker 40 seconds speech data are used for training and twice 3 seconds speech data are used for identifying.

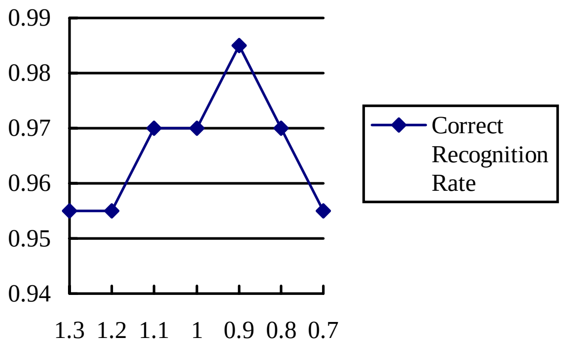


Figure 2: This graph illustrates the relationship between the correct recognition rate and the ITH.

As shown in the figure 2, When ITH is about 0.9, The performance of the system is best. It is much better than the system performance that not using this kind of data selection method (in this case ITH= ∞). This means that in speaker recognition not all frames of testing speech data have contributions to the final decision. Those speech data near a code-word's central vector are more useful.

3.OPEN-SET SPEAKER IDENTIFICATION SYSTEM

As mentioned above, the open-set speaker identification should give an additional decision alternative(whether the unknown is in the training set or not).

In the close-set speaker identification system, these two kinds of distance estimation methods (Absolute and Euclid distance estimation methods) get excellent performance. In our close-set system, the correct recognition rate is 98.5% when the training time is 40 seconds and the testing time is 10 seconds.

As mentioned above, the open-set speaker identification should give an additional decision alternative(whether the unknown is in the training set or not).

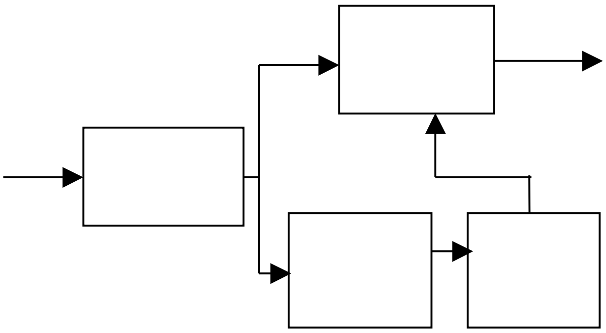
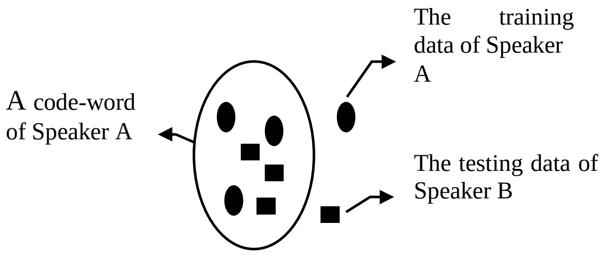


Figure 3: Block diagram of a VQ-Based open-set speaker identification system

In the close-set speaker identification system, these two kinds of distance estimation methods (Absolute and Euclid distance estimation methods) get excellent performance. In our close-set system, the correct recognition rate is 98.5% when the training time is 40 seconds and the testing time is 10 seconds.

In the open-set speaker identification system, we discover that the two formulae mentioned above are not applicable. Because the open-set identification system must give the rejection decision, It requires that the distance between every trainer's testing speech and his/her training speech must be smaller than the distance of most other speakers' testing speech and his/her own training speech. But the two formula mentioned above can not reach the requirement. The sketch map is as follows.



As shown above, in a code-word of Speaker A, there is apparent discrepancy between the speech data distribution of A and B. But the following situation will surely happen that the distance between testing speech data of speaker B and the code-word of speaker A is smaller than the distance between testing speech data of Speaker A and the code-word of Speaker A, whatever distance estimation formula (Absolute distance formula or Euclid distance formula) is adopted.

If the situation mentioned above emerges, the correct rejection decision can not be obtained. So, we propose another distance estimation method to estimate the distance between the testing speech data and the code-model. We call the method BCDM(Based on Codes Distribution Method).

In recognition, we firstly compute the distribution of testing speech data in the VQ code-model. That is, for every frame the distance between it and every code-word is computed. Then it can be attributed to the corresponding code-word according to the minimum distance. After processing all the frames, we can get the distribution in the VQ code-model of the testing speech data.

When computing the distance between the testing speech data and the VQ code-model, We adopt the following formula.

$$D_j = \frac{1}{M}\sum_{k=1}^{M}\left|F(P_k) - F\left(P_k^j\right)\right| \qquad (7)$$

F(*) is the probability distribution function of testing speech data in VQ code-model.

When we simply define:

$$F(x) = x \qquad (8)$$

We did the following experiment. The training set includes 40 speakers, 22 of them are males and 18 of them are females. There are altogether 20 persons out of the training set.

For each speaker in the training set 60 sec speech data is used for training and 18 sec is used for identifying. For each speaker out of the training set 18 sec is used for identifying.

The experiment result is:

The Equal Error Rate is 5%.

4.EXPERIMENT DATA

The speech data used in all experiments is recorded by computer in our laboratory. It is recorded by 16-bit Sound Blaster card and the sampling rate is 4KHZ. The interval between training and identifying is more than one month.

5. CONCLUSION

We have analyzed the classical VQ algorithm and made a lot of improvements. In the close-set speaker identification we adopt an advanced VQ algorithm, and introduce a new conception of the ITH. In the open-set speaker identification system we adopt BCDM(Based on Codes Distribution Method). As we make these improvements, the correct rate is fairly increased.

A lot of other work is needed to do in this speaker identification system. For example, we need to search a better distance estimation method and a better feature for the speaker identification problem[3].

6. REFERENCES

1. G.R.Doddington, ”Speaker Recognition-Identifying People by their Voices”. Proc. Institue of Acoustics, Vol.14.Part6, pp.95-100(1992).

2. T.Matusui and S.Furui,”A Text-Independent Speaker Recognition Method Robust Against Utterance Variations,” Proc.IEEE Int.Conf.Acoust.Speech Signal Processing,S6.3,pp.377-380(1991).

3. S.Furui “Recent Advances in Speaker Recognition”First International Conference, AVBPA’97,pp238-250.